特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2003 年7 月17 日 (17.07.2003)

PCT

(10) 国際公開番号
WO 03/058500 A1

(51) 国際特許分類[7]: G06F 17/30, C12Q 1/68, C12N 15/00, G01N 33/50

(21) 国際出願番号: PCT/JP03/00011

(22) 国際出願日: 2003 年1 月6 日 (06.01.2003)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願2001-402081
2001 年12 月28 日 (28.12.2001) JP

(71) 出願人 (米国を除く全ての指定国について): セレスター・レキシコ・サイエンシズ株式会社 (CELESTAR LEXICO-SCIENCES, INC.) [JP/JP]; 〒261-8501 千葉県 千葉市美浜区 中瀬１丁目３番地 幕張テクノガーデンＤ１７ Chiba (JP).

(72) 発明者; および
(75) 発明者/出願人 (米国についてのみ): 上村 泰央 (UEMURA,Yasuo) [JP/JP]; 〒261-8501 千葉県 千葉市美浜区 中瀬１丁目３番地 幕張テクノガーデンＤ１７ セレスター・レキシコ・サイエンシズ株式会社内 Chiba (JP). 蓬莱 尚幸 (HORAI,Hisayuki) [JP/JP]; 〒261-8501 千葉県 千葉市美浜区 中瀬１丁目３番地 幕張テクノガーデンＤ１７ セレスター・レキシコ・サイエンシズ株式会社内 Chiba (JP).

(74) 代理人: 酒井 宏明, 外 (SAKAI,Hiroaki et al.); 〒100-0013 東京都 千代田区 霞が関三丁目２番６号 東京倶楽部ビルディング Tokyo (JP).

(81) 指定国 (国内): US.

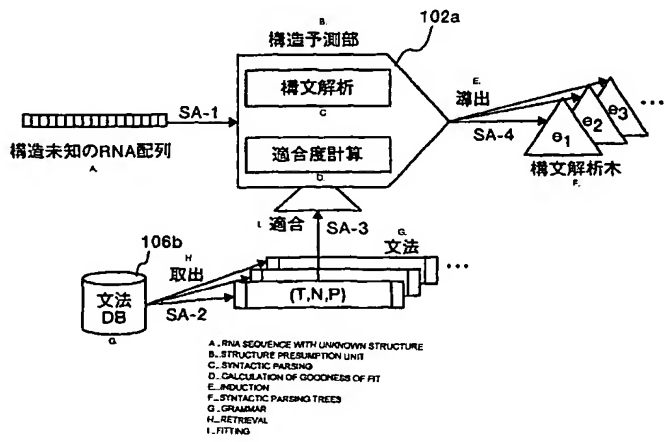
(84) 指定国 (広域): ヨーロッパ特許 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR).

添付公開書類:
— 国際調査報告書

２文字コード及び他の略語については、定期発行される各*PCT*ガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

(54) Title: RNA SEQUENCE ANALYZER, AND RNA SEQUENCE ANALYSIS METHOD, PROGRAM AND RECORDING MEDIUM

(54) 発明の名称: ＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体



(57) Abstract: An RNA sequence analyzer which is provided with a grammar storage unit wherein the structural topology of an RNA secondary structure and the corresponding generative grammar fitting for the topology are stored; a syntactic parsing unit wherein an RNA sequence is applied to the generative grammar to induce syntactic pausing trees; and a unit for calculating goodness of fit wherein the goodness of fit for the syntactic parsing trees induced by the syntactic parsing unit is calculated.

[続葉有]

WO 03/058500 A1

(57) 要約:

本発明にかかるＲＮＡ配列解析装置は、ＲＮＡ二次構造の構造トポロジーと当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段とを備えている。

明　細　書

ＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体

技術分野

本発明は、ＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体に関し、特に、ＲＮＡの二次構造を予測し、また、ＤＮＡ配列から遺伝子部分を予測するＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体に関する。

背景技術

ＲＮＡ配列は、Ａ（アデニン）、Ｃ（シトシン）、Ｇ（グアニン）、Ｕ（ウラシル）の４種の塩基により構成されるが、その一部は回文言語のような入れ子の状態となり、相補的な塩基同士（ＡとＵ、ＧとＣ、稀にＧとＵ）が結合して二次構造を構成している。ＲＮＡ配列の二次構造は、第１図に示すように、多種の構造トポロジーを有している。ここで、積み重ねられた塩基対の連続領域をステム（ｓｔｅｍ）と呼び、また、塩基対で挟まれた一本鎖の部分配列をループ（ｌｏｏｐ）と呼ぶ。ステムの端のループをヘアピンループという（第１図のａ）。ステムの中にある一本鎖の塩基は、それがステムの片側だけにあるときバルジループ（ｂｕｌｇｅ　ｌｏｏｐ）と呼び（第１図のｂ）、ステムの両側にあるときは内側ループ（ｉｎｔｅｒｎａｌ　ｌｏｏｐ）と呼ぶ（第１図のｃ）。放射状に３個以上のステムが出ているものをマルチループ（ｍｕｌｔｉ－ｂｒａｎｃｈｅｄ　ｌｏｏｐ）と呼ぶ。また、入れ子ではない塩基対があるとき、シュードノット（ｐｓｅｕｄｏｋｎｏｔｓ）と呼ぶ（第１図のｄ）。

ここで、ＲＮＡ配列を形式文法（生成文法）により構文解析することによりその二次構造を予測する手法が存在するが、正規文法では回文言語を記述することができないため、一般に、ＲＮＡ二次構造解析においては、木文法（ｔｒｅｅ

adjoining grammars）や、文脈自由文法（CFG）などを用いて構文解析を行い構造モデリング（構造トポロジー表現）を求める手法が考案されている。

例えば、Yasuo Uemura等著「Tree adjoining grammars for RNA structure prediction（Theoretical Computer Science 210 1999 277p－303p）」（以下「文献1」という）には、木文法による構造モデリングと、パーザ（paser）を利用したエネルギー極小化によるRNA二次構造予測方法が開示されている。

また、Elena Rivas and Sean R. Rddy著「The language of RNA: a formal grammar that includes pseudoknots（BIOINFORMATICS vol. 16 no. 4 2000 334p－340p）」（以下「文献2」という）には、Crossed－interaction Grammarsなどの独自の拡張を施した文脈自由文法（CFG）による構造モデリングと、パーザを利用したエネルギー極小化によるRNA二次構造予測方法が開示されている。

また、Michael Zuker著「Prediction of RNA Secondary Structure by Energy minimization（July 8, 1996）」（以下「文献3」という）には、形式文法やパーザを用いず動的計画法（Dynamic Programming）によってRNA二次構造を予測する方法を用いたRNA配列解析システムであるMfold（製品名）が開示されている。これらの文献では、形式文法や動的計画法などの手法と、エネルギー極小化手法とを組み合わせることによってRNA二次構造予測精度を高めている。

第2図は、従来技術によるRNAの二次構造がステムループをとる場合の構文解析木の一例を示す図である。第2図のaに示すRNA配列の二次構造を第2図

のｂに、また、構文解析木を第２図のｃに示している。ここで、部分木（ｓｕｂｔｒｅｅ）は、内部の節を根とする構文解析木の断片である。ＲＮＡ二次構造の構造トポロジーについて構文解析木を作成して構文解析を行うことにより二次構造解析を行う技術が研究されており、主要な構造トポロジーに対する文法が既知となっている。

第３図は、従来技術によるＲＮＡ二次構造の構造トポロジーについて、文法が固定されるとそれに対応した構造トポロジーが規定される（逆もまた可）ことを示す概念図である。ここで、生成文法（以下単に「文法」という）は、有限個の記号と、有限個の生成規則Ｐ（ｐｒｏｄｕｃｔｉｏｎ　ｒｕｌｅ）から成る。記号には、抽象的な非終端記号Ｎ（ｎｏｎｔｅｒｍｉｎａｌ　ｓｙｍｂｏｌ）と、観測文字列に実際に現れる終端記号Ｔ（ｔｅｒｍｉｎａｌ　ｓｙｍｂｏｌ）の２種類がある。終端記号Ｔは、ＲＮＡ配列の場合にはＡ、Ｔ、Ｇ、Ｃの４文字である。第３図に示すように、各構造トポロジーについてそれぞれ対応する文法を定義することができる。

また、第４図は、従来技術である木文法パーザを用いて、既知の文法からＲＮＡ配列の構文解析木を導出する場合の一例を示す図である。まず、構造未知のＲＮＡ配列を木文法パーザに入力する。ここで、木文法パーザは、入力された既知の木文法に従ってＲＮＡ配列の構文解析を行い構文解析木を導出する機能、および、導出された構文解析木について、ループや、塩基対とその他の二次構造要素のそれぞれの自由エネルギーの合計などを計算することにより平衡自由エネルギー（ΔＧ）などの値を求める機能などを有する（文献１から３参照）。

ここで、木文法パーザは必ずしも構文解析木を導出するわけではなく、入力したＲＮＡ配列が当該文法に適合しない場合（パーズが成功しない場合）には構文解析木を導出しない（すなわち、構文解析木は０個になる）。また、複数個の構文解析木が導出された場合には、エネルギー計算の結果、極小の自由エネルギーとなる１つの構文解析木を選択する。このとき、木文法パーザは導出過程の各段階において自由エネルギー極小な部分構造を見つけていくことができる。また、

木文法パーザはエネルギー準最適な構文も出力することができる。このように、木文法パーザは、構文解析（パーズ）の途中でエネルギー計算を行うことにより、高速化と精度向上を実現している。

しかしながら、従来の木文法パーザなどにより構文解析とエネルギー計算を行う手法を用いたＲＮＡ二次構造予測システムにおいては、ＲＮＡ配列や抽出した文法を統合的に管理し、集積した文法やＲＮＡ配列を用いて二次構造予測などをより効率的に行うものは存在しなかったという問題点がある。

また、与えられた特定の二次構造を取り得るようなＲＮＡ配列を検索するような方法は存在しなかったという問題点がある。

また、複数のＲＮＡ配列に共通な二次構造を簡易に抽出する方法は存在しなかったという問題点がある。

また、ＲＮＡ配列からＲＮＡ二次構造に基づく類似度を簡易に求める方法は存在しなかったという問題点がある。

さらに、ＤＮＡ配列から遺伝子部分を発見するための手法としては、ホモロジー検索やモチーフ検索などを利用する手法が一般的であるが、未知の遺伝子部分の発見には利用できないという問題点がある。ここで、従来技術で説明したように、ＲＮＡ配列の構造トポロジーを予測可能な生成文法が求められているが、既知の生成文法により導出された構文解析木を用いた遺伝子発見方法は存在しなかったという問題点がある。

このように、従来のシステム等は数々の問題点を有しており、その結果、システムの利用者および管理者のいずれにとっても、利便性が悪く、また、利用効率が悪いものであった。

従って、本発明は、ＲＮＡ配列や抽出した文法を統合的に管理し、集積した文法やＲＮＡ配列を用いて二次構造予測や新たな解析手法などをより効率的に行うことのできる、ＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体を提供することを目的としている。

発明の開示

本発明にかかるＲＮＡ配列解析装置は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソート手段と、上記ソート手段によりソートされた上記構文解析木を上記ＲＮＡ配列の二次構造の候補として出力する出力手段とを備えたことを特徴とする。

この装置によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算し、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートし、ソートされた構文解析木をＲＮＡ配列の二次構造の候補として出力するので、一配列に対して多文法で構文解析を行うことができるようになる。すなわち、各生成文法に対してそれぞれ構文解析し適合度計算を行い適合度を得る。その結果、生成文法ごとに適合度が得られることになり、それらの適合度をソートすることによって生成文法に順位を付ける。これにより、生成文法に対する構造トポロジーにも順位が付けられることになるので、最終的にＲＮＡ配列が取り得る可能性の高い順に構造トポロジーを確認することができるようになる。

つぎの発明にかかるＲＮＡ配列解析装置は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力する出力手段

とを備えたことを特徴とする。

この装置によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力するので、多配列に対して一文法で構文解析を行うことができるようになる。すなわち、与えられた特定の構造トポロジーに対し、対応する生成文法を取得し、これを用いてＲＮＡ配列データベースに格納されているすべてまたは一部のＲＮＡ配列をそれぞれ構文解析し、ある閾値以下の適合度で構文解析に成功したＲＮＡ配列群を結果として出力する。これにより、与えられた特定の二次構造を取り得るようなＲＮＡ配列を検索することができるようになる。

つぎの発明にかかるＲＮＡ配列解析装置は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出手段と、上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出手段にて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記ＲＮＡ配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成手段とを備えたことを特徴とする。

この装置によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲ

ＮＡ配列を抽出し、構造トポロジーとＲＮＡ配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたＲＮＡ配列と構造トポロジーに対応する格子部分をマークすることにより、ＲＮＡ配列間で共通に有する構造トポロジーを可視化するので、ＲＮＡ配列間の共通構造を容易に発見することができるようになる。

つぎの発明にかかるＲＮＡ配列解析装置は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成するＲＮＡ配列作成手段と、上記ＲＮＡ配列作成手段により作成された上記ＲＮＡ配列に対して上記生成文法を適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列に対応する上記ＤＮＡ配列部分を遺伝子の候補として予測する遺伝子予測手段とを備えたことを特徴とする。

この装置によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成し、作成されたＲＮＡ配列に対して生成文法を適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列に対応するＤＮＡ配列部分を遺伝子の候補として予測するので、ＤＮＡ配列のうち既知のトポロジーを有する可能性のあるＲＮＡ配列に対応する部分について、遺伝子部分である可能性があることを予測することができるようになる。

つぎの発明にかかるＲＮＡ配列解析装置は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析手段

と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度に基づいて上記ＲＮＡ配列間の類似度を計算する類似度計算手段とを備えたことを特徴とする。

この装置によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度に基づいてＲＮＡ配列間の類似度を計算するので、ＲＮＡ構造の類似度を容易に求めることができるようになる。

つぎの発明にかかるＲＮＡ配列解析装置は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出手段と、上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出手段にて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成手段と、上記適合度マトリックス作成手段にて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のＲＮＡ配列を抽出する共通構造抽出手段とを備えたことを特徴とする。

この装置によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構

文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を抽出し、構造トポロジーとＲＮＡ配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたＲＮＡ配列と構造トポロジーに対応する格子部分に適合度を表示する適合度マトリックスを作成し、適合度マトリックスについて、適合度により構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた構造トポロジーの順番に対応する生成文法により構文解析を行い適合度が最大となる構文解析木を求め、予め定めた条件を満たす適合度を持つ構文解析木に対応する他のＲＮＡ配列を抽出するので、共通の構造を持つＲＮＡ配列を容易に発見することができるようになる。

また、本発明はＲＮＡ配列解析方法に関するものであり、本発明にかかるＲＮＡ配列解析方法は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソートステップと、上記ソートステップによりソートされた上記構文解析木を上記ＲＮＡ配列の二次構造の候補として出力する出力ステップとを含むことを特徴とする。

この方法によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度を計算し、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートし、ソートされた構文解析木をＲＮＡ配列の二次構造の候補として出力するので、一配列に対して多文法で構文解析を行うことができるようになる。すなわち、各生成文法に対してそれぞれ構文解析し適合度計算を行い適合度を得る。

その結果、生成文法ごとに適合度が得られることになり、それらの適合度をソートすることによって生成文法に順位を付ける。これにより、生成文法に対する構造トポロジーにも順位が付けられることになるので、最終的にＲＮＡ配列が取り得る可能性の高い順に構造トポロジーを確認することができるようになる。

つぎの発明にかかるＲＮＡ配列解析方法は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力する出力ステップとを含むことを特徴とする。

この方法によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力するので、多配列に対して一文法で構文解析を行うことができるようになる。すなわち、与えられた特定の構造トポロジーに対し、対応する生成文法を取得し、これを用いてＲＮＡ配列データベースに格納されているすべてまたは一部のＲＮＡ配列をそれぞれ構文解析し、ある閾値以下の適合度で構文解析に成功したＲＮＡ配列群を結果として出力する。これにより、与えられた特定の二次構造を取り得るようなＲＮＡ配列を検索することができるようになる。

つぎの発明にかかるＲＮＡ配列解析方法は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合

度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出ステップと、上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記ＲＮＡ配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成ステップとを含むことを特徴とする。

この方法によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を抽出し、構造トポロジーとＲＮＡ配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたＲＮＡ配列と構造トポロジーに対応する格子部分をマークすることにより、ＲＮＡ配列間で共通に有する構造トポロジーを可視化するので、ＲＮＡ配列間の共通構造を容易に発見することができるようになる。

つぎの発明にかかるＲＮＡ配列解析方法は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成するＲＮＡ配列作成ステップと、上記ＲＮＡ配列作成ステップにより作成された上記ＲＮＡ配列に対して上記生成文法を適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列に対応する上記ＤＮＡ配列部分を遺伝子の候補として予測する遺伝子予測ステップとを含むことを特徴とする。

この方法によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジー

に適合する生成文法とを対応付けて格納し、利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成し、作成されたＲＮＡ配列に対して生成文法を適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列に対応するＤＮＡ配列部分を遺伝子の候補として予測するので、ＤＮＡ配列のうち既知のトポロジーを有する可能性のあるＲＮＡ配列に対応する部分について、遺伝子部分である可能性があることを予測することができるようになる。

つぎの発明にかかるＲＮＡ配列解析方法は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度に基づいて上記ＲＮＡ配列間の類似度を計算する類似度計算ステップとを含むことを特徴とする。

この方法によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度に基づいてＲＮＡ配列間の類似度を計算するので、ＲＮＡ構造の類似度を容易に求めることができるようになる。

つぎの発明にかかるＲＮＡ配列解析方法は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出ステップと、上記構造トポロジーと上記ＲＮ

A配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成ステップと、上記適合度マトリックス作成ステップにて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のＲＮＡ配列を抽出する共通構造抽出ステップとを含むことを特徴とする。

この方法によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を抽出し、構造トポロジーとＲＮＡ配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたＲＮＡ配列と構造トポロジーに対応する格子部分に適合度を表示する適合度マトリックスを作成し、適合度マトリックスについて、適合度により構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた構造トポロジーの順番に対応する生成文法により構文解析を行い適合度が最大となる構文解析木を求め、予め定めた条件を満たす適合度を持つ構文解析木に対応する他のＲＮＡ配列を抽出するので、共通の構造を持つＲＮＡ配列を容易に発見することができるようになる。

また、本発明はＲＮＡ配列解析方法をコンピュータに実行させるプログラムに関するものであり、本発明にかかるプログラムは、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算

された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソートステップと、上記ソートステップによりソートされた上記構文解析木を上記ＲＮＡ配列の二次構造の候補として出力する出力ステップとを含むことを特徴とする。

このプログラムによれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算し、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートし、ソートされた構文解析木をＲＮＡ配列の二次構造の候補として出力するので、一配列に対して多文法で構文解析を行うことができるようになる。すなわち、各生成文法に対してそれぞれ構文解析し適合度計算を行い適合度を得る。その結果、生成文法ごとに適合度が得られることになり、それらの適合度をソートすることによって生成文法に順位を付ける。これにより、生成文法に対する構造トポロジーにも順位が付けられることになるので、最終的にＲＮＡ配列が取り得る可能性の高い順に構造トポロジーを確認することができるようになる。

つぎの発明にかかるプログラムは、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力する出力ステップとを含むことを特徴とする。

このプログラムによれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、

計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力するので、多配列に対して一文法で構文解析を行うことができるようになる。すなわち、与えられた特定の構造トポロジーに対し、対応する生成文法を取得し、これを用いてＲＮＡ配列データベースに格納されているすべてまたは一部のＲＮＡ配列をそれぞれ構文解析し、ある閾値以下の適合度で構文解析に成功したＲＮＡ配列群を結果として出力する。これにより、与えられた特定の二次構造を取り得るようなＲＮＡ配列を検索することができるようになる。

つぎの発明にかかるプログラムは、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出ステップと、上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記ＲＮＡ配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成ステップとを含むことを特徴とする。

このプログラムによれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を抽出し、構造トポロジーとＲＮＡ配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたＲＮＡ配列と構造トポロジーに対応する格子部分をマークすることにより、ＲＮＡ配列間で共通に有する構造トポロジーを可視化するので、ＲＮＡ配列間の共通構造を容易に発見することが

できるようになる。

つぎの発明にかかるプログラムは、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成するＲＮＡ配列作成ステップと、上記ＲＮＡ配列作成ステップにより作成された上記ＲＮＡ配列に対して上記生成文法を適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列に対応する上記ＤＮＡ配列部分を遺伝子の候補として予測する遺伝子予測ステップとを含むことを特徴とする。

このプログラムによれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成し、作成されたＲＮＡ配列に対して生成文法を適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列に対応するＤＮＡ配列部分を遺伝子の候補として予測するので、ＤＮＡ配列のうち既知のトポロジーを有する可能性のあるＲＮＡ配列に対応する部分について、遺伝子部分である可能性があることを予測することができるようになる。

つぎの発明にかかるプログラムは、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度に基づいて上記ＲＮＡ配列間の類似度を計算する類似度計算ステップとを含むことを特徴とする。

このプログラムによれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度に基づいてＲＮＡ配列間の類似度を計算するので、ＲＮＡ構造の類似度を容易に求めることができるようになる。

つぎの発明にかかるプログラムは、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出ステップと、上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成ステップと、上記適合度マトリックス作成ステップにて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のＲＮＡ配列を抽出する共通構造抽出ステップとを含むことを特徴とする。

このプログラムによれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を抽出し、構造トポロジーとＲＮＡ配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたＲＮＡ配列と構造トポロジー

に対応する格子部分に適合度を表示する適合度マトリックスを作成し、適合度マトリックスについて、適合度により構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた構造トポロジーの順番に対応する生成文法により構文解析を行い適合度が最大となる構文解析木を求め、予め定めた条件を満たす適合度を持つ構文解析木に対応する他のＲＮＡ配列を抽出するので、共通の構造を持つＲＮＡ配列を容易に発見することができるようになる。

また、本発明は記録媒体に関するものであり、本発明にかかる記録媒体は、上記に記載されたプログラムを記録したことを特徴とする。

この記録媒体によれば、当該記録媒体に記録されたプログラムをコンピュータに読み取らせて実行することによって、上記に記載されたプログラムをコンピュータを利用して実現することができ、これら各プログラムと同様の効果を得ることができる。

図面の簡単な説明

第１図は、ＲＮＡの構造トポロジーの一例を説明する図であり、第２図は、従来技術によるＲＮＡの二次構造がステムループをとる場合の構文解析木の一例を示す図であり、第３図は、従来技術によるＲＮＡ二次構造の構造トポロジーについて、文法が固定されるとそれに対応した構造トポロジーが規定されることを示す概念図であり、第４図は、従来技術である木文法パーザを用いて、既知の文法からＲＮＡ配列の構文解析木を導出する場合の一例を示す図であり、第５図は、本発明が適用される本システムの構成の一例を示すブロック図であり、第６図は、文法データベース１０６ｂに格納される情報の一例を示す図であり、第７図は、本実施の形態における本システムのＲＮＡ二次構造予測処理の一例を示す処理概念図であり、第８図は、本実施の形態における本システムの同一構造ＲＮＡ配列抽出処理の一例を示す処理概念図であり、第９図は、本実施の形態における本システムの共通構造抽出処理の一例を示す処理概念図であり、第１０図は、本実施の形態における本システムの構造類似度計算処理の一例を示す処理概念図であり、

第１１図は、本実施の形態における本システムの遺伝子予測処理の一例を示す処理概念図であり、第１２図は、ペナルティＰと類似度ベクトルs_1、s_2の概念を説明する図であり、第１３図は、ＲＮＡ二次構造トポロジーの例を示す図であり、第１４図は、s_1の構文解析木と二次構造を示す図であり、第１５図は、塩基対の自由エネルギーを示す図であり、第１６図は、ループの自由エネルギーを示す図であり、第１７図は、それぞれの文法について－ΔＧの適合度指標において最適な構文解析木とそれに対応する二次構造を示す図であり、第１８図は、選択されたトポロジー集合のなかでs_2が適合する構造候補を示す図であり、第１９図は、選択されたトポロジーをとりうる配列の候補を示す図であり、第２０図は、構文解析木の適合度を要素に持つマトリックスを示す図であり、第２１図は、ｓの最適な二次構造を示す図であり、第２２図は、構文解析木の適合度を要素に持つマトリックスを示す図であり、第２３図は、出力結果の一例を示す図である。

発明を実施するための最良の形態

以下に、本発明にかかるＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体の実施の形態を図面に基づいて詳細に説明する。なお、この実施の形態によりこの発明が限定されるものではない。

特に、以下の実施の形態においては、本発明を、木文法に適用した例について説明するが、この場合に限られず、全ての生成文法において、同様に適用することができる。

[本システムの概要]

以下、本システムの概要について説明し、その後、本システムの構成および処理等について詳細に説明する。

このシステムは、概略的に、以下の基本的特徴を有する。すなわち、本システムのＲＮＡ配列解析装置は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度を計算し、計算

された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートし、ソートされた構文解析木をＲＮＡ配列の二次構造の候補として出力する。ここで、生成文法は、木文法、文脈自由文法などを含むが、シュードノットを表現するためには木文法が最も適しているため、木文法を用いることが好ましい。

また、本装置は、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力する。

また、本装置は、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を抽出し、構造トポロジーとＲＮＡ配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたＲＮＡ配列と構造トポロジーに対応する格子部分をマークすることにより、ＲＮＡ配列間で共通に有する構造トポロジーを可視化する。

また、本装置は、利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成し、作成されたＲＮＡ配列に対して生成文法を適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列に対応するＤＮＡ配列部分を遺伝子の候補として予測する。

さらに、本装置は、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度に基づいてＲＮＡ配列間の類似度を計算する。

［システム構成］

まず、本システムの構成について説明する。第５図は、本発明が適用される本システムの構成の一例を示すブロック図であり、該構成のうち本発明に関係する部分のみを概念的に示している。本システムは、概略的に、配列情報を解析する

ＲＮＡ配列解析装置であるＲＮＡ配列解析装置１００と、配列情報等に関する外部データベースやホモロジー検索用の外部分析プログラム等を提供する外部システム２００とを、ネットワーク３００を介して通信可能に接続して構成されている。

第５図においてネットワーク３００は、ＲＮＡ配列解析装置１００と外部システム２００とを相互に接続する機能を有し、例えば、インターネット等である。

第５図において外部システム２００は、ネットワーク３００を介して、ＲＮＡ配列解析装置１００と相互に接続され、利用者に対して配列情報等に関する外部データベースやホモロジー検索やモチーフ検索等の外部分析プログラムを実行するウェブサイトを提供する機能を有する。

ここで、外部システム２００は、ＷＥＢサーバやＡＳＰサーバ等として構成してもよく、そのハードウェア構成は、一般に市販されるワークステーション、パーソナルコンピュータ等の情報処理装置およびその付属装置により構成してもよい。また、外部システム２００の各機能は、外部システム２００のハードウェア構成中のＣＰＵ、ディスク装置、メモリ装置、入力装置、出力装置、通信制御装置等およびそれらを制御するプログラム等により実現される。

第５図においてＲＮＡ配列解析装置１００は、概略的に、ＲＮＡ配列解析装置１００の全体を統括的に制御するＣＰＵ等の制御部１０２、通信回線等に接続されるルータ等の通信装置（図示せず）に接続される通信制御インターフェース部１０４、入力装置１１２および出力装置１１４に接続される入出力制御インターフェース部１０８、および、各種のデータベースやテーブル（ＲＮＡ配列データベース１０６ａ～共通構造マトリックス１０６ｃ）を格納する記憶部１０６を備えて構成されており、これら各部は任意の通信路を介して通信可能に接続されている。さらに、このＲＮＡ配列解析装置１００は、ルータ等の通信装置および専用線等の有線または無線の通信回線を介して、ネットワーク３００に通信可能に接続されている。

記憶部１０６に格納される各種のデータベース（ＲＮＡ配列データベース１０

６ａ～共通構造マトリックス１０６ｃ）は、固定ディスク装置等のストレージ手段であり、各種処理に用いる各種のプログラムやテーブルやファイルやデータベースやウェブページ用ファイル等を格納する。

これら記憶部１０６の各構成要素のうち、ＲＮＡ配列データベース１０６ａは、ＲＮＡ配列を格納したデータベースである。ＲＮＡ配列データベース１０６ａは、インターネットを経由してアクセスする外部のＲＮＡ配列データベースであってもよく、また、これらのデータベースをコピーしたり、オリジナルの配列情報を格納したり、さらに独自のアノテーション情報等を付加したりして作成したインハウスデータベースであってもよい。また、ＲＮＡ配列データベース１０６ａは、ｃＤＮＡ等のＤＮＡ配列データベースに基づいて予め生成された、あるいは必要時に動的に生成されたＲＮＡ配列を格納したものでもよい。

また、文法データベース１０６ｂは、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段である。ここで、第６図は、文法データベース１０６ｂに格納される情報の一例を示す図である。第６図に示すように、文法データベース１０６ｂは、構造トポロジーと、その構造トポロジーに対応する文法とを対応付けて格納する。ここで、文法データベース１０６ｂには、第６図に示したように、構造トポロジーと文法とが１対１で対応するようにしてもよく、また、複数のトポロジーが結合した文法（例えば、シュードノットとヘアピンループとを両方持つトポロジーなど）や、特徴的な構造を有するＲＮＡ用の文法（例えば、ｒＲＮＡに特有の構造トポロジーなど）や、所定のカテゴリーのＲＮＡが共通で備えるトポロジー用の文法や、全てのＲＮＡに適合する文法などを規定してもよい。

また、共通構造マトリックス１０６ｃは、構造トポロジーとＲＮＡ配列とを二次元マトリックスで表示するためのテーブル（記憶領域）である。

また、第５図において、通信制御インターフェース部１０４は、ＲＮＡ配列解析装置１００とネットワーク３００（またはルータ等の通信装置）との間における通信制御を行う。すなわち、通信制御インターフェース部１０４は、他の端末

と通信回線を介してデータを通信する機能を有する。

また、第５図において、入出力制御インターフェース部１０８は、入力装置１１２や出力装置１１４の制御を行う。ここで、出力装置１１４としては、モニタ（家庭用テレビを含む）の他、スピーカを用いることができる（なお、以下においては出力装置をモニタとして記載する）。また、入力装置１１２としては、キーボード、マウス、および、マイク等を用いることができる。また、モニタも、マウスと協働してポインティングデバイス機能を実現する。

また、第５図において、制御部１０２は、ＯＳ（Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）等の制御プログラム、各種の処理手順等を規定したプログラム、および所要データを格納するための内部メモリを有し、これらのプログラム等により、種々の処理を実行するための情報処理を行う。制御部１０２は、機能概念的に、構造予測部１０２ａ、類似度計算部１０２ｄ、共通構造マトリックス作成部１０２ｆ、および、遺伝子予測部１０２ｇを備えて構成されている。

このうち、構造予測部１０２ａは、入力された既知の文法に従ってＲＮＡ配列の構文解析を行い構文解析木を導出する機能（構文解析部１０２ｂ）、および、導出された構文解析木に対して適合度の計算を行う機能（適合度計算部１０２ｃ）などを有する。

また、類似度計算部１０２ｄは、複数のＲＮＡ配列間の類似度を計算する類似度計算手段である。

また、共通構造マトリックス作成部１０２ｆは、適合度計算手段により計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を抽出する抽出手段、および、構造トポロジーとＲＮＡ配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出手段にて抽出されたＲＮＡ配列と構造トポロジーに対応する格子部分をマークすることにより、ＲＮＡ配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成手段、二次元マトリックスにおいて抽出手段にて抽出されたＲＮＡ配列と構造トポロジーに対応する格子部分に適合度を表示する適合度マトリックスを作成する適

合度マトリックス作成手段、および、適合度マトリックス作成手段にて作成された適合度マトリックスについて、適合度により構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた構造トポロジーの順番に対応する生成文法により構文解析を行い適合度が最大となる構文解析木を求め、予め定めた条件を満たす適合度を持つ構文解析木に対応する他のＲＮＡ配列を抽出する共通構造抽出手段である。

また、遺伝子予測部１０２ｇは、利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成するＲＮＡ配列作成手段、および、適合度計算手段により計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列に対応するＤＮＡ配列部分を遺伝子の候補として予測する遺伝子予測手段である。なお、これら各部によって行なわれる処理の詳細については、後述する。

［システムの処理］

次に、このように構成された本実施の形態における本システムの処理の一例について、以下に第７図～第１１図を参照して詳細に説明する。

［ＲＮＡ二次構造予測処理］

まず、ＲＮＡ二次構造予測処理の詳細について第７図を参照して説明する。第７図は、本実施の形態における本システムのＲＮＡ二次構造予測処理の一例を示す処理概念図である。

まず、文法データベース１０６ｂに既知のＲＮＡの構造トポロジーを表す文法を集積する。そして、利用者が構造未知のＲＮＡ配列であってその二次構造を特定したいものを入力装置１１２を介してＲＮＡ配列解析装置１００に入力すると（ステップＳＡ－１）、構造予測部１０２ａは、構文解析部１０２ｂの処理により、文法データベース１０６ｂから文法を取り出し（ステップＳＡ－２）、ＲＮＡ配列に対して各文法を適合して構文解析（パーズ）を行う（ステップＳＡ－３）。ここで、利用者のＲＮＡ配列の入力は、ＲＮＡ配列データベース１０６ａから所望の配列を選択することにより入力してもよく、外部システム２００の外部

データベースから所望の配列を選択することにより入力してもよく、さらに、所望の配列を直接入力してもよい。

そして、構造予測部１０２ａは、適合度計算部１０２ｃの処理により、パーズが成功し導出された構文解析木について、例えば、ループや、塩基対とその他の二次構造要素のそれぞれの自由エネルギーの合計などを計算することにより求める平衡自由エネルギー（ΔＧ）などに基づいて適合度を求める。ここで、適合度計算方法は、例えば上述した文献１から３に示した方法のほか、従来のいずれの方法を用いてもよい。

そして、構造予測部１０２ａは、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートする（ステップＳＡ－４）。

そして、構造予測部１０２ａは、入出力制御インターフェース部１０８を介して出力装置１１４にソートされた構文解析木とその適合度を出力することにより、利用者が入力した一配列に対して多文法で構文解析を行うことができるようになる。すなわち、各生成文法に対してそれぞれ構文解析し適合度計算を行い適合度を得る。その結果、生成文法ごとに適合度が得られることになり、それらの適合度をソートすることによって生成文法に順位を付ける。これにより、生成文法に対する構造トポロジーにも順位が付けられることになるので、最終的にＲＮＡ配列が取り得る可能性の高い順に構造トポロジーを確認することができるようになる。これにて、ＲＮＡ二次構造予測処理が終了する。

［同一構造ＲＮＡ配列抽出処理］

次に、同一構造ＲＮＡ配列抽出処理の詳細について第８図を参照して説明する。第８図は、本実施の形態における本システムの同一構造ＲＮＡ配列抽出処理の一例を示す処理概念図である。

まず、利用者は、特定の構造トポロジーに対応する文法を文法データベース１０６ｂから選択する。そして、構造予測部１０２ａは、構文解析部１０２ｂの処理により、ＲＮＡ配列データベース１０６ａからＲＮＡ配列を取り出し（ステッ

プＳＢ－１）、各ＲＮＡ配列に対して文法を適合して（ステップＳＢ－２）、構文解析を行う（ステップＳＢ－３）。

そして、適合度計算部１０２ｃは、導出された構文解析木に対して適合度計算を行う。そして、構造予測部１０２ａは、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を、指定した文法が表す構造トポロジーの二次構造を持つＲＮＡ配列の候補として抽出する（ステップＳＢ－４）。

そして、構造予測部１０２ａは、抽出されたＲＮＡ配列を当該文法が表す構造トポロジーの二次構造をもつ可能性のあるものとして、入出力制御インターフェース部１０８を介して出力装置１１４に出力する（ステップＳＢ－５）。これにて、同一構造ＲＮＡ配列抽出処理が終了する。

［共通構造抽出処理］

次に、共通構造抽出処理の詳細について第９図を参照して説明する。第９図は、本実施の形態における本システムの共通構造抽出処理の一例を示す処理概念図である。

まず、構造予測部１０２ａは、ＲＮＡ配列データベース１０６ａから１つまたは２つ以上のＲＮＡ配列を取り出し（ステップＳＣ－１およびステップＳＣ－２）、構文解析部１０２ｂの処理により、各ＲＮＡ配列に対して、文法データベース１０６ｂから取り出した（ステップＳＣ－３）、１つまたは２つ以上の文法を適合する（ステップＳＣ－４）。ＲＮＡ配列解析装置１００は、これらの取り出しや、パーズ処理について並列処理を行ってもよく、また、順次処理を行ってもよい。

そして、適合度計算部１０２ｃは、導出された構文解析木に対して適合度計算を行い、共通構造マトリックス作成部１０２ｆの処理により、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を抽出する（ステップＳＣ－５）。

そして、共通構造マトリックス作成部１０２ｆは、適合した文法が表す構造ト

ポロジーとＲＮＡ配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたＲＮＡ配列と構造トポロジーに対応する格子部分をマークすることにより、ＲＮＡ配列間で共通に有する構造トポロジーを可視化する（ステップＳＣ－６）。

ここで、マークは、第９図に示すように対象の格子部分に特定の色をつけてもよく、また、特定の記号（○など）や文字（「Ｙ」など）を記載してもよい。これにより、例えば、縦方向にマークが連続した場合（第９図に示す例においては、２番目の構造トポロジーの列）には、この構造トポロジーが各ＲＮＡ配列に共通に持っている配列であることが可視的に把握することができるようになる。これにて、共通構造抽出処理が終了する。

［構造類似度計算処理］

次に、構造類似度計算処理の詳細について第１０図を参照して説明する。第１０図は、本実施の形態における本システムの構造類似度計算処理の一例を示す処理概念図である。

まず、利用者が、入力装置１１２により類似度を計算したい複数（第１０図の例では２個）のＲＮＡ配列をＲＮＡ配列解析装置１００に入力する（ステップＳＥ－１）。

そして、類似度計算部１０２ｄは、文法データベース１０６ｂから１つまたは２つ以上の文法を取り出し（ステップＳＥ－２）、構文解析部１０２ｂの処理により、入力したＲＮＡ配列について文法を適合して構文解析を行う（ステップＳＥ－３）。また、適合度計算部１０２ｃは、導出された構文解析木に対して適合度計算を行う（ステップＳＥ－４）。

そして、類似度計算部１０２ｄは、文法を適合することにより導出された構文解析木と適合度（導出されなかった場合には、それを表現する特別な値を設定する）を各ＲＮＡ配列ごとに対応付けてベクトル演算や内積の計算などをすることにより（ステップＳＥ－５）、ＲＮＡ配列間の類似度を計算する（ステップＳＥ－６）。

例えば、入力であるｉ個のＲＮＡ配列をRNA_1，RNA_2，．．．，RNA_iとし、文法データベース１０６ｂに格納されているＮ個の文法をG_1，G_2，．．．，G_Nとし、また、ＲＮＡ配列ｘと文法ｇに対するパーザが成功したときの適合度をｒ（ｘ，ｇ）とする。ここで、適合度は実数値とし、適合度が大きいほどその構造を取り易いことを示すものとする。

また、入力RNA_jに関する適合度のベクトルR_jにおいて、R_jのｋ番目の要素$R_j[k]$は、RNA_jとG_kに対するパーザが成功したときはｒ（RNA_j，G_k）とし、RNA_jとG_kに対するパーザが失敗したときは仮に「×」とする。

このとき、類似度計算部１０２ｄによる類似度計算は、以下の手法により行われる。まず、２つのＲＮＡ配列の適合度のベクトルR_1とR_2を入力する。

ついで、類似度計算部１０２ｄは、類似度ベクトルS_1、S_2とペナルティＰを求める。ここで、「ペナルティＰ」は、$R_1[k]$と$R_2[k]$の片方だけが「パーザ失敗（×）」であるｋの個数であり、「類似度ベクトルS_1、S_2」は、$R_1[k]$も$R_2[k]$も「パーザ失敗（×）」ではない箇所だけを抜き出したベクトルである。第１２図は、ペナルティＰと類似度ベクトルS_1、S_2の概念を説明する図である。

ついで、類似度計算部１０２ｄは、類似度ベクトルS_1、S_2間の距離Ｄを以下の方法により求める。まず、類似度ベクトルS_1、S_2の要素数（ベクトルの次元）をＭとする。そして、類似度計算で一般的に用いられるユークリッド距離を用いて以下の数式により距離を計算する。

$$D = sqrt(\Sigma \{ (S_1[k] - S_2[k])^2 \})$$

（ｓｑｒｔは平方根であり、Σはｋ＝１～Ｍに関する総和である。）

ここで、距離Ｄが大きい場合には類似度が低いことになり、また、ペナルティＰが大きいと類似度が低いことになるので、ペナルティＰと距離Ｄを用いて以下の数式により類似度Ｓｉｍを求める。

$Sim = a^P / D$

（aは定数（$0 < a < 1$）である。）

そして、Ｓｉｍを類似度として出力する。ここで、定数ａを小さくすると、距離Ｄよりもペナルティ Ｐが重視されることになる。これにて、構造類似度計算処理が終了する。

［遺伝子予測処理］

次に、遺伝子予測処理の詳細について第１１図を参照して説明する。第１１図は、本実施の形態における本システムの遺伝子予測処理の一例を示す処理概念図である。

まず、利用者が遺伝子部分が未知のＤＮＡ配列を入力装置１１２を介してＲＮＡ配列解析装置１００に入力すると、ＲＮＡ配列解析装置１００は、遺伝子予測部１０２ｇの処理により、入力されたＤＮＡ配列に基づいて、当該ＤＮＡ配列から転写されるＲＮＡ配列（以下、「予測ＲＮＡ配列」という）を自動的に変換して作成する（ステップＳＦ－１）。ここで、利用者のＤＮＡ配列の入力は、外部システム２００の外部データベースやインハウスデータベースから所望のＤＮＡ配列を選択することにより入力してもよく、さらに、所望の配列を直接入力してもよい。

ついで、構造予測部１０２ａがこの予測ＲＮＡ配列を構文解析部１０２ｂに入力すると（ステップＳＦ－２）、構文解析部１０２ｂの処理により、文法データベース１０６ｂから１つまたは２つ以上の文法が取り出され（ステップＳＦ－３）、各文法を予測ＲＮＡ配列に適合する（ステップＳＦ－４）。

そして、適合度計算部１０２ｃは、構文解析部１０２ｂが導出した構文解析木について適合度計算を行い（ステップＳＦ－５）、遺伝子予測部１０２ｇは、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出した予測ＲＮＡ配列に対応するＤＮＡ配列部分を遺伝子の候補として予測する（ス

テップＳＦ－６）。すなわち、ＤＮＡ配列のうち、当該予測ＲＮＡ配列の部分が遺伝子部分である可能性が高い領域として出力される。

これにより、ＤＮＡ配列のうち既知のトポロジーを有する可能性のある予測ＲＮＡ配列に対応する部分について、遺伝子部分である可能性があることを予測することができるようになる。これにて、遺伝子予測処理が終了する。

［実施例］

本発明の実施例について、以下に第１３図～第２３図を参照して説明する。

１　準備

本節では実施例の準備として、いくつかの具体的なＲＮＡ二次構造トポロジーを定義し、それらをモデリングする生成文法を規定する。本実施例では説明の便宜上、生成文法として文脈自由文法を用いるが、よりモデリング能力の高いＲＮＡ木文法（文献１）を用いた場合でも同様のことが説明できる。

１．１　　二次構造トポロジー

第１３図に示される２つのＲＮＡ二次構造トポロジーを考える。

ステムループは、ステム（Ｈ（ａ））とヘアピンループ（Ｌ（ａ））から構成される。２並列ステムループは、並列に並んだ２つのステムループから構成される。それぞれのステム部分（H_1（ｂ）、H_2（ｂ））とヘアピンループ部分、（L_1（ｂ）、L_2（ｂ））の他にステムとステムをつなぐループ部分（Ｉ（ｂ））がある。

上記構造トポロジーについて、さらに具体的な特徴を考えることができる。例えばステムやループ部分のサイズの制約、ステムを構成する塩基対にミスマッチ（内部ループやバルジループ）を許すかどうか、あるいは特定の場所に特定の塩基配列を含むかどうか、などといったより詳細な特徴を持ったトポロジーを考えることが可能である。そこで、本実施例では次のような特徴を持ったＲＮＡ二次構造トポロジーT_1、T_2を扱う。

トポロジーT_1

－　　以下の特徴を持ったステムループ構造（第１３図（ａ）参照）である。

－　　ステム（H（a））を構成する塩基対はミスマッチを含まない。

－　　ステム（H（a））サイズは１塩基対以上とする。

－　　ヘアピンループ（L（a））のサイズは１塩基以上とする。

トポロジーT_2

－　　以下の特徴を持った２並列ステムループ構造（第１３図（b）参照）である。

－　　トポロジーT_1を２つ並列に並べたもの。

－　　ステム（H_1（b））とステム（H_2（b））の間のループ（I（b））の長さは１塩基以上とする。

１．２　　文脈自由文法による二次構造トポロジーのモデリング

以上のように定義された２つのトポロジーT_1、T_2を文脈自由文法を用いてモデリングする。文脈自由文法は一般に以下の４項組によって規定される。

$G=(N, \Sigma, P, S)$

Nは非終端記号の有限集合、Σは終端記号の有限集合、Pは生成規則の有限集合、Sは開始記号を表す。

しかしながら、本実施例では常に$\Sigma=\{a, u, g, c\}$、開始記号はS、さらにNは生成規則Pに出現する非終端記号のみからなるものとすることから、Pのみを指定することにより文脈自由文法Gを規定することができる。よって便宜上、本稿では文脈自由文法Gを規定する際には、生成規則の有限集合Pのみを指定することにする。

（１）トポロジーT_1は以下の生成規則からなる文脈自由文法G_1によってモデリングされる。

$$S \rightarrow xH\bar{x}$$
$$H \rightarrow xH\bar{x} \mid L$$
$$L \rightarrow xL \mid x$$

ただし、$x \in \Sigma$とし、$\bar{x}$はxと塩基対を形成する相手の塩基とする。

すなわち、ワトソンークリック塩基対のみを考える場合は、一番最初の生成規則は以下と同値である。

S→aHu | uHa | gHc | cHg

非ワトソンークリック塩基対を許す場合はさらに、S→gHuなどを追加してもよい。

G_1において、

$$S \rightarrow xH\bar{x} \text{ と } H \rightarrow xH\bar{x}$$

によって塩基対（ステムを構成）が生成され、L→xLとL→xによって塩基対を形成しない塩基（ループを構成）が生成されるとみなす。すると、G_1はRNAの二次構造を生成することができることになる。このように、任意の文脈自由文法Gに対して、Gが生成することができるすべてのRNA二次構造からなる集合SS（G）が規定されることになる。

G_1がトポロジーT_1をモデリングする、とは以下が成り立つことを言う。“G_1はトポロジーT_1の性質を満たすすべてのRNA二次構造を生成することができ、なおかつ、G_1で生成することができるすべてのRNA二次構造はトポロジーT_1の性質を満たす。”

これは、G_1による導出をみれば自明である。G_1による導出はすべて以下のよ

うになる。

ただし、$n \geq 1$、$l \geq 1$とする。

$$
\begin{aligned}
&S \rightarrow x_1 H \overline{x_1} \\
&\rightarrow x_1 x_2 H \overline{x_2 x_1} \rightarrow \ldots \rightarrow x_1 x_2 \ldots x_n H \overline{x_n} \ldots \overline{x_2 x_1} \\
&\rightarrow x_1 x_2 \ldots x_n L \overline{x_n} \ldots \overline{x_2 x_1} \\
&\rightarrow x_1 x_2 \ldots x_n y_1 L \overline{x_n} \ldots \overline{x_2 x_1} \\
&\rightarrow \ldots \rightarrow x_1 x_2 \ldots x_n y_1 \ldots y_{l-1} L \overline{x_n} \ldots \overline{x_2 x_1} \\
&\rightarrow x_1 x_2 \ldots x_n y_1 \ldots y_{l-1} y_l \overline{x_n} \ldots \overline{x_2 x_1}
\end{aligned}
$$

このとき、$x_1 x_2 \ldots x_n$と、これに対応する$\overline{x_n} \ldots \overline{x_2 x_1}$の部分がステムに、$y_1 \ldots y_{l-1} y_l$の部分がヘアピンループに対応する。また、$n \geq 1$、$l \geq 1$であるからステムのサイズは1塩基対以上、ヘアピンループのサイズは1塩基以上である。

よって、G_1はT_1をモデリングすることがわかる。

（2）トポロジーT_2は以下の生成規則からなる文脈自由文法G_2によってモデリングされる。

$$
\begin{aligned}
&S \rightarrow S_1 I S_2 \\
&S_1 \rightarrow x H \overline{x} \\
&S_2 \rightarrow x H \overline{x} \\
&H \rightarrow x H \overline{x} \mid L \\
&L \rightarrow x L \mid x \\
&I \rightarrow x L
\end{aligned}
$$

以下の生成規則からなる文脈自由文法G_0は、文脈自由文法によって生成することが可能なすべてのＲＮＡ二次構造を生成することができる万能な文脈自由文法である。

$$S \rightarrow SS \mid xS\overline{x} \mid xS \mid Sx \mid x \mid \lambda$$

ただし、λ は空文字を表す。例えば、G_1によるいかなる導出もG_0によってシミュレートできる。すなわち、以下のような導出をG_0によって行なうことが可能である。

$$\begin{aligned}
S &\rightarrow x_1 S \overline{x_1} \\
&\rightarrow x_1 x_2 S \overline{x_2 x_1} \rightarrow \ldots \rightarrow x_1 x_2 \ldots x_n S \overline{x_n} \ldots \overline{x_2 x_1} \\
&\rightarrow x_1 x_2 \ldots x_n S \overline{x_n} \ldots \overline{x_2 x_1} \\
&\rightarrow x_1 x_2 \ldots x_n y_1 S \overline{x_n} \ldots \overline{x_2 x_1} \\
&\rightarrow \ldots \rightarrow x_1 x_2 \ldots x_n y_1 \ldots y_{l-1} S \overline{x_n} \ldots \overline{x_2 x_1} \\
&\rightarrow x_1 x_2 \ldots x_n y_1 \ldots y_{l-1} y_l \overline{x_n} \ldots \overline{x_2 x_1}
\end{aligned}$$

上記の導出は非終端記号以外、すなわち生成されるＲＮＡ二次構造は、G_1によって生成されるものとまったく同じである。よってG_1が生成可能なすべての二次構造をG_0によって生成可能なことがわかる。すなわち、

$$SS(G_0) \supseteq SS(G_1)$$

である。

このように、どんな文脈自由文法Ｇに対しても

$$SS(G_0) \supseteq SS(G)$$

が成り立つことが知られている。以降では、このような万能文法によって生成される二次構造全体を“すべての二次構造”と考える。

１．３　構文解析木と適合度

ある与えられたＲＮＡ配列がある与えられたＲＮＡ二次構造トポロジーの性質を満たす二次構造を形成できるかどうかという問題は、対象トポロジーをモデリングした文法によって対象配列が導出できるかという問題に対応する。これは生成文法の構文解析アルゴリズムによって解くことができる。

構文解析アルゴリズムは、与えられた文法によって、与えられた配列が導出できるかどうかを判定し、導出可能な場合はその導出過程、すなわち構文解析木を出力する。二次構造トポロジーをモデリングした文法において、構文解析木は二次構造を表現しているので、構文解析アルゴリズムは、対象トポロジーに適合する具体的な二次構造を出力すると解釈してよいことになるからである。

ＲＮＡ配列s_1＝ｇｇｇｇａａａｃｃｃｃがトポロジーT_1、T_2に適合する二次構造を形成できるかどうかについて考える。

配列s_1はG_1によって以下のように導出できる。これにより配列s_1はT_1に適合する二次構造をとりうることがわかる。

Ｓ→ｇＨｃ→ｇｇＨｃｃ→ｇｇｇＨｃｃｃ→ｇｇｇｇＨｃｃｃｃ
→ｇｇｇｇＬｃｃｃｃ→ｇｇｇｇａＬｃｃｃｃ→ｇｇｇｇａａＬｃｃｃｃ
→ｇｇｇｇａａａｃｃｃｃ　　(1)

また、s_1はG_1によって以下のようにも導出できる。

Ｓ→ｇＨｃ→ｇｇＨｃｃ→ｇｇｇＨｃｃｃ
→ｇｇｇＬｃｃｃ→ｇｇｇｇＬｃｃｃ→ｇｇｇｇａＬｃｃｃ
→ｇｇｇｇａａＬｃｃｃ→ｇｇｇｇａａａＬｃｃｃ
→ｇｇｇｇａａａｃｃｃｃ　　(2)

しかし、s_1はG_2によって導出することはできない。これによりs_1はトポロ

ジーT_2に適合する二次構造をとりえないことがわかる。

s_1をG_1によって上記の２通りの方法で導出するとき、それぞれの導出に対応する構文解析木とそれが表現する二次構造を第１４図に示す。すなわち、（１）のように導出した場合は、第１４図の（１）に示される構文解析木と二次構造が生成され、（２）のように導出した場合は、第１４図の（２）に示される構文解析木と二次構造が生成される。

この例のように複数の構文解析木が得られたときに、どの構文解析木、すなわち、どの二次構造を結果として出力するかを決定する必要が生じる。そのために、何らかの評価関数によって構文解析木（あるいは二次構造）にスコアを与え、構文解析木（あるいは二次構造）に順位を付ける必要が生じる。このようなスコアは文法によって異なる評価関数を用いても良いし、文法に依存しない絶対的な評価関数であってもよい。以降では、このスコアを適合度と呼ぶ。

以下に、これまでに利用されている適合度の評価法の例を示すが、本発明によって利用される適合度は以下のものに限定されない。

（１）　塩基対数による適合度の評価

一般に、塩基対を形成する際の水素結合によってＲＮＡ分子はエネルギー的に安定なものになる。そこでこの評価法では、単純に塩基対の数が多い二次構造ほど優先させる。つまり、構文解析木の適合度として、対応する二次構造の塩基対数を用いる。この評価法において、上記の例の適合度を評価すると、第１４図の（１）に示される構文解析木は適合度３となり、（２）に示される構文解析木は適合度２となり、適合度の大きい（１）の構造が採用されることになる。

本評価法に基づいた代表的な手法として、Ｎｕｓｓｉｎｏｖの折り畳みアルゴリズム［Ｎｕｓｓｉｎｏｖ，Ｒ．，Ｐｉｅｃｘｅｎｋ，Ｇ．，ｇｅｉｇｇｓ，ｊ．Ｒ．，ａｎｄ　Ｋｌｅｉｔｍａｎ，Ｄ．Ｊ．，“Ａｌｇｏｒｉｔｈｍｓ　ｆｏｒ　ｌｏｏｐ　ｍａｔｃｈｉｎｇｓ，”　ＳＩＡＭ　ｊｏｕｒｎａｌ　ｏｆ　Ａｐｐｌｉｅｄ　Ｍａｔｈｅｍａｔｉｃｓ，　３５，　６８－８２，１９７８］がある。

（２）　平衡自由エネルギー（ΔＧ）による適合度の評価

ＲＮＡ二次構造の物理化学的な安定度を計算するために、小さなモデルＲＮＡ分子の熱力学的な実験によって決定された平衡自由エネルギー（ΔＧ）パラメータを利用する方法がある。ある二次構造の（ΔＧ）は、それを構成する塩基対やループなどの二次構造要素に対する自由エネルギーの合計で近似される。この自由エネルギーパラメータでは、塩基対により構造が安定化し、ループにより構造が不安定化する。各二次構造要素の詳細なパラメータは［Ｔｕｒｎｅｒ，　Ｄ．Ｈ．，Ｓｕｇｉｍｏｔｏ，Ｎ．，Ｊａｅｇｅｒ，Ｊ．Ａ．，Ｌｏｎｇｆｅｌｌｏｗ，Ｃ．Ｅ．，Ｆｒｅｉｅｒ，Ｓ．Ｍ．，ａｎｄ　Ｋｉｅｒｚｅｋ，Ｒ．，“Ｉｍｐｒｏｖｅｄ　ｐａｒａｍｅｔｅｒｓ　ｆｏｒ　ｐｒｅｄｉｃｔｉｏｎ　ｏｆ　ＲＮＡ　ｓｔｒｕｃｔｕｒｅ，”　Ｃｏｌｄ　Ｓｐｒｉｎｇ　Ｈａｒｂｏｒ　Ｓｙｍｐｏｓｉａ　Ｑｕａｎｔｉｔａｔｉｖｅ　Ｂｉｏｌｏｇｙ，　５２，１２３－１３３，１９８７］に示されている。ここでは塩基対の自由エネルギーを第１５図に、ループの自由エネルギーを第１６図に示す。

上記の自由エネルギーパラメータを用いて第１４図の構造（１）と（２）の（ΔＧ）を求めると、それぞれ以下のようになる。

ΔＧ（構造（１））＝ΔＧ（ｇｃ，ｇｃ）＋ΔＧ（ｇｃ，ｇｃ）
＋ΔＧ（ｇｃ，ｇｃ）
＋（ΔＧ）（サイズ３のヘアピンループ）
＝（－２．９）＋（－２．９）＋（－２．９）
＋７．４＝－１．３

ΔＧ（構造（２））＝ΔＧ（ｇｃ，ｇｃ）＋ΔＧ（ｇｃ，ｇｃ）
＋ΔＧ（サイズ５のヘアピンループ）
＝（－２．９）＋（－２．９）＋４．４＝－１．４

ここで注意すべきことは、塩基対の自由エネルギーの計算方法である。連続して積み重なった２組の塩基対に対してひとつのエネルギー値が与えられる。すな

わち、構造（１）では５’側から数えて１番目のｇｃ塩基対と２番目のｇｃ塩基対に対して、ΔＧ（ｇｃ，ｇｃ）が計算され、２番目のｇｃ塩基対と３番目のｇｃ塩基対に対して、ΔＧ（ｇｃ，ｇｃ）が計算され、３番目のｇｃ塩基対と４番目のｇｃ塩基対に対して、ΔＧ（ｇｃ，ｇｃ）が計算される。これに対し構造（２）では５’側から数えて１番目のｇｃ塩基対と２番目のｇｃ塩基対に対して、ΔＧ（ｇｃ，ｇｃ）が計算され、２番目のｇｃ塩基対と３番目のｇｃ塩基対に対して、ΔＧ（ｇｃ，ｇｃ）が計算される。

構文解析木の適合度を－ΔＧと定めると、（１）の適合度は１．３となり、（２）の適合度は１．４となり、結果として適合度の大きい（２）の構造が採用されることになる。

ΔＧに基づいた代表的なＲＮＡ二次構造予測システムとして、ＺｕｋｅｒのＭｆｏｌｄ（文献３）がある。

（３）　導出確率による適合度の評価

確率文法とは個々の生成規則にその適用確率が付加された生成文法である。例えばG_1の各生成規則に以下のような確率ｐが付加されている確率文脈自由文法G_1を考える。

ｐ（Ｓ→ａＨｕ）＝０．２

ｐ（Ｓ→ｕＨａ）＝０．２

ｐ（Ｓ→ｇＨｃ）＝０．３

ｐ（Ｓ→ｃＨｇ）＝０．３

ｐ（Ｈ→ａＨｕ）＝０．２

ｐ（Ｈ→ｕＨａ）＝０．２

ｐ（Ｈ→ｇＨｃ）＝０．３

ｐ（Ｈ→ｃＨｇ）＝０．２

ｐ（Ｈ→Ｌ）＝０．１

ｐ（Ｌ→ａＬ）＝０．２

p（L→uL）＝0．2

p（L→gL）＝0．15

p（L→cL）＝0．15

p（L→a）＝0．1

p（L→u）＝0．1

p（L→g）＝0．05

p（L→c）＝0．05

このときG_1によるs_1の導出確率は次のようにして計算される。すなわち、（1）の導出確率は、

p（S→gHc）×p（H→gHc）×p（H→gHc）×p（H→gHc）×p（H→L）×p（L→aL）×p（L→aL）×p（L→a）

＝0．3×0．3×0．3×0．3×0．1×0．2×0．2×0．1

＝0．00000324

と計算される。また、（2）の導出確率は、

p（S→gHc）×p（H→gHc）×p（H→gHc）×p（H→L）×p（L→gL）×p（L→aL）×p（L→aL）×p（L→aL）×p（L→c）

＝0．3×0．3×0．3×0．1×0．15×0．2×0．2×0．2

×0．05

＝0．000000162

となる。

そこで構文解析木の適合度として導出確率の自然対数をとると、（1）の適合度は1n0．00000324＝－12．6、（2）の適合度は1n0．000

０００１６２＝－１５．６となり、結果として適合度の大きい（１）の構造が採用されることになる。

本評価法の根本である、各生成規則に付加されるべき確率パラメータは、最尤推定法と内側・外側アルゴリズム（ｉｎｓｉｄｅ－ｏｕｔｓｉｄｅ ａｌｇｏｒｉｔｈｍ）などにより学習してもよいし、ヒューリスティクス（ｈｅｕｒｉｓｔｉｃｓ）などによって主観的に推定してもよい。例えば、文献［Ｓａｋａｋｉｂａｒａら“Ｓｔｏｃｈａｓｔｉｃ Ｃｏｎｔｅｘｔ－ｆｅｅ Ｇｒａｍｍａｒｓ ｆｏｒ ｔＲＮＡ ｍｏｄｅｌｉｎｇ，” Ｎｕｃｌｅｉｃ Ａｃ.ｉｄｓ Ｒｅｓｅａｒｃｈ， ２２，５１１２－５１２０，１９９４．］では複数のｔＲＮＡ配列からｔＲＮＡの構造をモデリングする確率文脈自由文法を学習する手法について述べられている。

以上いくつかの適合度評価法について説明したが、以降の説明では適合度として－ΔＧを用いる。

次にＲＮＡ配列s_2＝ｇｃｃｃａｕａｇｇｃａａａｇｃｃｕａｕｇｇｇｃがトポロジーT_1、T_2に適合する二次構造を形成できるかどうかを考える。この場合も同様にs_2がG_1、G_2によって導出できるかどうかを調べればよい。結論からいうと、s_2はG_1、G_2のいずれからも導出することができる。さらにどちらの文法でも複数の導出のしかたが存在する。それぞれの文法について－ΔＧの適合度指標において最適な構文解析木とそれに対応する二次構造を第１７図に示す。

それぞれの構造のΔＧを求めると、以下のようになる。

ΔＧ（構造（１））＝ΔＧ（ｇｃ，ｃｇ）×２＋ΔＧ（ｃｇ，ｃｇ）
×２＋ΔＧ（ｃｇ，ａｕ）＋ΔＧ（ａｕ，ｕａ）
＋ΔＧ（ｕａ，ａｕ）＋ΔＧ（ａｕ，ｇｃ）
＋ΔＧ（ｇｃ，ｇｃ）
＋ΔＧ（サイズ３のヘアピンループ）
＝（－３．４）×２＋（－２．９）×２＋（－１．８）

＋（－０．９）＋（－１．１）＋（－１．７）
＋（－２．９）＋７．４＝－１３．６

ΔＧ（構造（２））＝ΔＧ（ｇｃ，ｃｇ）×２＋ΔＧ（ｃｇ，ｃｇ）×２
＋ΔＧ（サイズ４のヘアピンループ）×２
＝（－３．４）×２＋（－２．９）×２＋５．９×２
＝－６．７

よってトポロジーT_1に適合するＲＮＡ二次構造においてs_2がとりうる最適な構造の適合度は１３．６であることがわかる。また、トポロジーT_2に適合するＲＮＡ二次構造においてs_2がとりうる最適な構造の適合度は６．７であることがわかる。また、s_2を万能な文法G_0によって構文解析すると、最適構造として構造（１）が見つかる。これにより、構造（１）が“すべての二次構造”の中で最適な構造であることがわかる。このように万能文法によって構文解析を行なうことにより、すべての構造の中から最適な構造を見つけ出すことができる。

本発明の基本となる“ＲＮＡ配列を生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された構文解析木に対して適合度の計算を行う適合度計算手段と、適合度最大の構文解析木に対応する二次構造を出力する最適二次構造出力手段”は、適合度計算を組み込んだ構文解析アルゴリズムにより実現されることになる。このような構文解析アルゴリズムを構造予測アルゴリズムと呼ぶ。ΔＧの適合度を指標にしたＲＮＡ木文法に基づく構造予測アルゴリズムは文献１に示されている。

２．　発明実施例

本節では、前節までに定義したＲＮＡ配列s_1、s_2、トポロジーT_1、T_2およびそれらをモデリングする文脈自由文法G_0、G_1、G_2、さらに適合度として－ΔＧを用いた場合の実施例について示す。

はじめに、“ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合

する生成文法とを対応付けて格納する文法格納手段”においては、例えば（Ｌｅｕ－ｔＲＮＡ，Ｇ’）や（１６Ｓ　ｒＲＮＡ，Ｇ”）などのようにある構造トポロジーに付けられた名称とその構造トポロジーをモデリングした文法が対応づけられて格納されている。本実施例では（ステムループT_1，G_1）と（２並列ステムループT_2，G_2）を含むような文法ＤＢを仮定する。また、ＲＮＡ配列s_1とs_2を含むＲＮＡ配列ＤＢを仮定する。

（１）生成文法と適合度計算で構造候補を出力

あるＲＮＡ配列に対して、これがとりうる構造トポロジーを適合度が大きい順に知りたいとき、本発明によれば、以下の手順でこれを調べることができる。例として、入力配列をs_2、検査対象トポロジー集合をT_1、T_2とした場合について示す。

手順１）　ＲＮＡ配列を配列ＤＢから指定する。あるいは直接入力する。ここではs_2を指定する。

手順２）　検査対象とするトポロジーの集合（生成文法の集合）を文法ＤＢから選択する。ここではT_1とT_2（G_1とG_2）を選択する。

手順３）　適合度の閾値を設定する。閾値は手順２で得られた各トポロジー（生成文法）に対してそれぞれ設定してもよいし、共通にひとつ設定してもよい。ここではT_1（G_1）に対して１０をT_2（G_2）に対して５を設定する。

手順４）　手順１で得られた配列を手順２で得られた各生成文法によってそれぞれ構文解析を行ない、適合度最大の構文解析木を求める。ここではs_2をG_1によって構文解析し、最大の適合度１３．６を持つ構文解析木を得る（第１７図（１）参照）。

さらにs_2をG_2によって構文解析し、最大の適合度６．７を持つ構文解析木を得る（第１７図（２）参照）。

手順５）　手順４で得られた構文解析木のうち手順３で得られた閾値以上の適合度をもつ構文解析木を適合度の大きい順にソートする。手順４で得られたG_1に対する適合度１３．６の構文解析木１は手順３でG_1に対して設定された閾値１

０よりも大きいのでこれをソート対象とする。手順４で得られたG_2に対する適合度６．７の構文解析木２は手順３でG_2に対して設定された閾値５よりも大きいのでこれをソート対象とする。上でソート対象となった構文解析木を適合度の大きい順にソートすることによって、構文解析木１、構文解析木２の順に順序づけされる。

手順６）　手順５でソートされた構文解析木の順に、対応するトポロジー名、適合度、構文解析木（二次構造）などを出力する。構文解析木１に対応して、ステムループT_1、適合度１３．６、第１７図（１）に示された二次構造を出力する。構文解析木２に対応して、２並列ステムループT_2、適合度６．７、第１７図（２）に示された二次構造を出力する。

以上の結果から、選択されたトポロジー集合のなかでs_2が適合する構造候補が第１８図のように出力される。

従来の二次構造予測プログラムでは、与えられた配列がとりうる構造のなかで最適あるいは準最適な二次構造を順に出力するだけで、出力された構造がどのようなトポロジーであるかはユーザが判断しなければならなかった。本発明によれば、構造とトポロジーとを対応付けて出力することができるので、予測結果の確認にかかる労力が大きく軽減されることが期待される。

また、本発明の実施について厳密に上記の手順と同じである必要はない。例えば、手順１と２は順序をいれかえてもよいし、手順５の閾値により構文解析木を取捨選択する部分は手順４の構文解析部分に含めてしまってもよい。

（２）同じ構造を持つ配列の候補を出力

ある構造トポロジーに対して、これに適合する二次構造をとりうるＲＮＡ配列を探したいとき、本発明によれば、以下の手順でこれを調べることができる。例として、入力構造トポロジーをT_2、検査対象配列集合をs_1、s_2とした場合について示す。

手順１）　トポロジー（生成文法）を文法ＤＢから選択する。ここではT_2（G_2）を選択する。

手順２）　適合度の閾値を設定する。ここでは５を選択する。

手順３）　検査対象とするＲＮＡ配列集合を配列ＤＢから選択する。あるいは直接入力する。ここではs_1、s_2選択する。

手順４）　手順３で得られた各配列を、手順１で得られた生成文法によってそれぞれ構文解析を行ない、適合度最大の構文解析木をそれぞれ求める。ここではs_1をG_2によって構文解析し、導出不能であることを得る。さらにs_2をG_2によって構文解析し、最大の適合度６．７を持つ構文解析木を得る。（第１７図（２）参照）

手順５）　手順４で得られた構文解析木のうち手順２で得られた閾値以上の適合度をもつ構文解析木に対応する配列を出力する。手順４で得られたs_2のG_2に対する適合度６．７の構文解析木は手順２で設定された閾値５よりも大きいのでs_2を出力する。以上の結果から、選択されたトポロジーをとりうる配列の候補が第１９図のように出力される。

本発明の実施について厳密に上記の手順と同じである必要はない。例えば、手順１と２と３は任意の順にいれかえてもよいし、手順５は手順４の構文解析部分に含めてしまってもよい。

（３）共通構造抽出

あるＲＮＡ配列の集合に対して、これらの配列が共通してとりうる構造トポロジーを調べたいとき、本発明によれば、以下の手順でこれを調べることができる。例として、入力配列集合をs_1、s_2とし、検査対象構造トポロジーの集合をT_1、T_2とした場合について示す。

手順１）　ＲＮＡ配列の集合を配列ＤＢから指定する。あるいは直接入力する。ここではs_1とs_2を指定する。

手順２）　検査対象とするトポロジーの集合（生成文法の集合）を文法ＤＢから選択する。ここではT_1（G_1）とT_2（G_2）を選択する。

手順３）　適合度の閾値を設定する。閾値は手順２で得られた各トポロジー（生成文法）に対してそれぞれ設定してもよいし、共通にひとつ設定してもよい。こ

こでは共通に０を設定する。

手順４）　手順１で得られた各配列を、手順２で得られた各生成文法によってそれぞれ構文解析を行ない、適合度最大の構文解析木を求める。

s_1をG_1によって構文解析し、最大の適合度１．４を持つ構文解析木を得る（第１４図（２）参照）。

s_1をG_2によって構文解析し、導出不能であることを得る。

s_2をG_1によって構文解析し、最大の適合度１３．６を持つ構文解析木を得る。（第１７図（１）参照）

s_2をG_2によって構文解析し、最大の適合度６．７を持つ構文解析木を得る。（第１７図（２）参照）

手順５）　手順４で得られた構文解析木のうち閾値以上の適合度を持つ構文解析木を抽出する。手順４で得られたすべての構文解析木は手順３で得られた閾値０よりも大きい適合度を持つので手順４で得られたすべての構文解析木を抽出する。

手順６）　手順１で得られた配列集合を行に、手順２で得られたトポロジー集合を列に、手順５で得られた構文解析木の適合度を要素に持つマトリックスを作成する。第２０図に示すマトリックスを得る。

以上の結果得られたマトリックスを出力すれば、対象配列集合が共通してとりうる構造トポロジーを容易に確認することが可能になる。あるいは、以下の追加手順を実行すれば、共通構造の候補を順位づけて出力することができる。

手順７）　手順６で得られたマトリックスの各列、すなわちトポロジー、に対してスコアを計算する。例えば、有効な行要素の数を各列ごとに計算しスコアとすると、T_1のスコアは２、T_2のスコアは１となる。例えば、各行の適合度の総和を各列ごとに計算しスコアとすると、T_1のスコアは１５．０、T_2のスコアは６．７となる。

手順８）　手順７で得られたスコアの高い順にトポロジーをソートし、出力する。上記のいずれのスコアを用いてもT_1、T_2の順に出力される。

また、本発明の実施について厳密に上記の手順と同じである必要はない。例え

ば、手順１と２は順序をいれかえてもよいし、手順５を手順４の構文解析部分に含めてしまってもよい。

（４）　ジーン・ファインダ

ＲＮＡ遺伝子部分に対応する配列は、非常に安定な構造をとりやすいので、適合度が高くなる。そこで本発明では、万能文法を用いて構文解析を行ない、適合度の高い配列を配列ＤＢから選び出して遺伝子候補として出力する。例として、配列集合をs_1、s_2とした場合について示す。

手順１）　検査対象とするＲＮＡ配列の集合を配列ＤＢから指定する。あるいは直接入力する。ここではs_1とs_2を指定する。

手順２）　適合度の閾値を設定する。ここでは１０を設定する。

手順３）　手順１で得られた各配列を万能文法G_0によってそれぞれ構文解析を行ない、適合度最大の構文解析木を求める。

s_1をG_0によって構文解析し、最大の適合度１．４を持つ構文解析木を得る。

s_2をG_0によって構文解析し、最大の適合度１３．６を持つ構文解析木を得る。

手順４）　手順３で得られた構文解析木のうち閾値以上の適合度をもつ構文解析木に対応する配列を遺伝子候補として出力する。手順３で得られたs_1の構文解析木は閾値１０に満たないのでs_1は出力しない。手順３で得られたs_2の構文解析木は閾値１０よりも大きいのでs_2を遺伝子候補として出力する。

本発明の実施について厳密に上記の手順と同じである必要はない。例えば、手順１と２は順序をいれかえてもよいし、手順４は手順３の構文解析部分に含めてしまってもよい。

（５）　ＲＮＡ配列から同じ構造を持つＲＮＡ配列を出力

あるＲＮＡ配列集合に対して、これらと同じトポロジーをとりうるＲＮＡ配列を調べたいとき、（３）の発明と（２）の発明とを組み合わせた本発明によれば、以下の手順でこれを調べることができる。例として、入力配列をｓ＝ｇｃｃｃａａａａｇｇｇｃａｇｃｃｃａａａｇｇｇｃ、検査対象トポロジー集合をT_1、T_2、検査対象配列集合をs_1、s_2とした場合について示す。

手順１）　ＲＮＡ配列集合を入力する。ここではｓのみからなる配列集合を入力する。

手順２）　検査対象とするＲＮＡ配列の集合を配列ＤＢから指定する。ここではs_1とs_2を指定する。

手順３）　検査対象とするトポロジーの集合（生成文法の集合）を文法ＤＢから選択する。ここではT_1（G_1）とT_2（G_2）を選択する。

手順４）　適合度の閾値を設定する。閾値は手順３で得られた各トポロジー（生成文法）に対してそれぞれ設定してもよいし、共通にひとつ設定してもよい。ここでは共通に５を設定する。

手順５）　手順１で得られた各ＲＮＡ配列を、手順２で得られた各生成文法によってそれぞれ構文解析を行ない、適合度最大の構文解析木をそれぞれ求める。ここではｓをG_1によって構文解析し、最大の適合度３．１を持つ構文解析木を得る。第２１図（１）にこの構文解析木が表現する二次構造を示す。さらにｓをG_2によって構文解析し、最大の適合度５．１を持つ構文解析木を得る。第２１図（２）にこの構文解析木が表現する二次構造を示す。

手順６）　手順５で得られた構文解析木のうち、手順４で得られた閾値以上の適合度をもつ構文解析木に対応する構文解析木を抽出する。手順５で得られた構文解析木のうち、G_2で構文解析することによって得られた適合度５．１の構文解析木が手順４で得られた閾値５よりも大きいのでこれを抽出する。

手順７）　手順１で得られた配列集合を行に、手順３で得られたトポロジー集合を列に、手順６で得られた構文解析木の適合度を要素に持つマトリックスを作成する。第２２図に示すマトリックスを得る。

手順８）　手順６で得られたマトリックスの各列、すなわちトポロジー、に対してスコアを計算し、スコアの順にトポロジーをソートする。ここでは行の総和を各列ごとに計算しスコアとするが、１行しかないので結果として、T_1のスコアが未定義、T_2のスコアが５．１になる。スコアを持つものだけでソートすると、T_2のみが得られる。

手順９）　手順２で得られた各配列を、手順８で得られたトポロジーの順にそれぞれ対応する文法で構文解析を行ない、適合度最大の構文解析木をそれぞれ求める。ここではs_1をG_2によって構文解析し、導出不能であることを得る。

さらにs_2をG_2によって構文解析し、最大の適合度６．７を持つ構文解析木を得る。（第１７図（２）参照）

手順１０）　手順９で得られた構文解析木のうち手順４で得られた閾値以上の適合度をもつ構文解析木に対応する配列を出力する。このとき、あわせてトポロジーとそのトポロジーに対する手順８で得られたスコアを出力する。手順９で得られたs_2のG_2に対する構文解析木の適合度６．７は手順４で得られた閾値５よりも大きいのでs_2を出力する。あわせて、T_2とそのスコア５．１を出力する。

以上の結果から第２３図に示すような出力が得られる。

この結果、s_2がトポロジーT_2において、ｓと共通な構造をとりうることがわかるようになる。

本発明の実施について厳密に上記の手順と同じである必要はない。例えば、手順１と２と３は任意の順に入れ換えてもよいし、手順６は手順５の構文解析部分に含めてしまってもよいし、手順１０の閾値により構文解析木を取捨選択する部分は手順９の構文解析部分に含めてしまってもよい。

[他の実施の形態]

さて、これまで本発明の実施の形態について説明したが、本発明は、上述した実施の形態以外にも、上記特許請求の範囲に記載した技術的思想の範囲内において種々の異なる実施の形態にて実施されてよいものである。

例えば、ＲＮＡ配列解析装置１００がスタンドアローンの形態でＲＮＡ配列解析方法を行う場合を一例に説明したが、ＲＮＡ配列解析装置１００とは別筐体で構成されるクライアント端末からの要求に応じてＲＮＡ配列解析方法を行い、その処理結果を当該クライアント端末に返却するように構成してもよい。

また、構造予測部１０２ａは、適合度計算部１０２ｃによる適合度計算を行いながら構文解析部１０２ｂにより構文解析木を導出してもよい。すなわち、構文

解析木を導出する構文解析部１０２ｂと、導出された構文解析木に対して適合度の計算を行う適合度計算部１０２ｃをひとつのアルゴリズムにて実現してもよい。このように構成することにより、ＲＮＡ配列と木文法に対して可能な構文解析木は無数（配列長に対して指数のオーダー）に存在するため、構文解析木を導出してから適合度計算を行いソートすると指数オーダーの計算時間と記憶容量が必要となるという問題点を解決することができる。

また、実施の形態において説明した各処理のうち、自動的に行なわれるものとして説明した処理の全部または一部を手動的に行うこともでき、あるいは、手動的に行なわれるものとして説明した処理の全部または一部を公知の方法で自動的に行うこともできる。

特に、構造予測部１０２ａは複数のタスクとして実現してもよく、それぞれのタスクで並列処理を行うように実現してもよい。

この他、上記文書中や図面中で示した処理手順、制御手順、具体的名称、各種の登録データや検索条件等のパラメータを含む情報、画面例、データベース構成については、特記する場合を除いて任意に変更することができる。

また、ＲＮＡ配列解析装置１００に関して、図示の各構成要素は機能概念的なものであり、必ずしも物理的に図示の如く構成されていることを要しない。

例えば、ＲＮＡ配列解析装置１００の各サーバが備える処理機能、特に制御部にて行なわれる各処理機能については、その全部または任意の一部を、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）および当該ＣＰＵにて解釈実行されるプログラムにて実現することができ、あるいは、ワイヤードロジックによるハードウェアとして実現することも可能である。なお、プログラムは、後述する記録媒体に記録されており、必要に応じてＲＮＡ配列解析装置１００に機械的に読み取られる。

記憶部１０６に格納される各種のデータベース等（ＲＮＡ配列データベース１０６ａ～共通構造マトリックス１０６ｃ）は、ＲＡＭ、ＲＯＭ等のメモリ装置、ハードディスク等の固定ディスク装置、フレキシブルディスク、光ディスク等の

ストレージ手段であり、各種処理やウェブサイト提供に用いる各種のプログラムやテーブルやファイルやデータベースやウェブページ用ファイル等を格納する。

また、ＲＮＡ配列解析装置１００は、既知のパーソナルコンピュータ、ワークステーション等の情報処理端末等の情報処理装置にプリンタやモニタやイメージスキャナ等の周辺装置を接続し、該情報処理装置に本発明の方法を実現させるソフトウェア（プログラム、データ等を含む）を実装することにより実現してもよい。

さらに、ＲＮＡ配列解析装置１００の分散・統合の具体的形態は図示のものに限られず、その全部または一部を、各種の負荷等に応じた任意の単位で、機能的または物理的に分散・統合して構成することができる。例えば、各データベースを独立したデータベース装置として独立に構成してもよく、また、処理の一部をＣＧＩ（Ｃｏｍｍｏｎ　Ｇａｔｅｗａｙ　Ｉｎｔｅｒｆａｃｅ）を用いて実現してもよい。

また、本発明にかかるプログラムを、コンピュータ読み取り可能な記録媒体に格納することもできる。ここで、この「記録媒体」とは、フレキシブルディスク、光磁気ディスク、ＲＯＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭ、ＣＤ－ＲＯＭ、ＭＯ、ＤＶＤ等の任意の「可搬用の物理媒体」や、各種コンピュータシステムに内蔵されるＲＯＭ、ＲＡＭ、ＨＤ等の任意の「固定用の物理媒体」、あるいは、ＬＡＮ、ＷＡＮ、インターネットに代表されるネットワークを介してプログラムを送信する場合の通信回線や搬送波のように、短期にプログラムを保持する「通信媒体」を含むものとする。

また、「プログラム」とは、任意の言語や記述方法にて記述されたデータ処理方法であり、ソースコードやバイナリコード等の形式を問わない。なお、「プログラム」は必ずしも単一的に構成されるものに限られず、複数のモジュールやライブラリとして分散構成されるものや、ＯＳ（Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）に代表される別個のプログラムと協働してその機能を達成するものをも含む。なお、実施の形態に示した各装置において記録媒体を読み取るための具体的な構

成、読み取り手順、あるいは、読み取り後のインストール手順等については、周知の構成や手順を用いることができる。

また、ネットワーク３００は、ＲＮＡ配列解析装置１００と外部システム２００とを相互に接続する機能を有し、例えば、インターネットや、イントラネットや、ＬＡＮ（有線／無線の双方を含む）や、ＶＡＮや、パソコン通信網や、公衆電話網（アナログ／デジタルの双方を含む）や、専用回線網（アナログ／デジタルの双方を含む）や、ＣＡＴＶ網や、ＩＭＴ２０００方式、ＧＳＭ方式またはＰＤＣ／ＰＤＣ―Ｐ方式等の携帯回線交換網／携帯パケット交換網や、無線呼出網や、Ｂｌｕｅｔｏｏｔｈ等の局所無線網や、ＰＨＳ網や、ＣＳ、ＢＳまたはＩＳＤＢ等の衛星通信網等のうちいずれかを含んでもよい。すなわち、本システムは、有線・無線を問わず任意のネットワークを介して、各種データを送受信することができる。

以上詳細に説明したように、本発明によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度を計算し、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートし、ソートされた構文解析木をＲＮＡ配列の二次構造の候補として出力するので、一配列に対して多文法で構文解析を行うことができるようになる。すなわち、各生成文法に対してそれぞれ構文解析し適合度計算を行い適合度を得る。その結果、生成文法ごとに適合度が得られることになり、それらの適合度をソートすることによって生成文法に順位を付ける。これにより、生成文法に対する構造トポロジーにも順位が付けられることになるので、最終的にＲＮＡ配列が取り得る可能性の高い順に構造トポロジーを確認することができるＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体を提供することができる。

また、本発明によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用し

て構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力するので、多配列に対して一文法で構文解析を行うことができるようになる。すなわち、与えられた特定の構造トポロジーに対し、対応する生成文法を取得し、これを用いてＲＮＡ配列データベースに格納されているすべてまたは一部のＲＮＡ配列をそれぞれ構文解析し、ある閾値以下の適合度で構文解析に成功したＲＮＡ配列群を結果として出力する。これにより、与えられた特定の二次構造を取り得るようなＲＮＡ配列を検索することができるＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体を提供することができる。

また、本発明によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を抽出し、構造トポロジーとＲＮＡ配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたＲＮＡ配列と構造トポロジーに対応する格子部分をマークすることにより、ＲＮＡ配列間で共通に有する構造トポロジーを可視化するので、ＲＮＡ配列間の共通構造を容易に発見することができるＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体を提供することができる。

また、本発明によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成し、作成されたＲＮＡ配列に対して生成文法を適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列に対応するＤＮＡ配列部分を遺伝子の候補として予測するので、ＤＮＡ配列のうち既知のトポロジーを有する可能性のあるＲＮＡ配列に対応する

部分について、遺伝子部分である可能性があることを予測することができるＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体を提供することができる。

また、本発明によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度に基づいてＲＮＡ配列間の類似度を計算するので、ＲＮＡ構造の類似度を容易に求めることができるＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体を提供することができる。

さらに、本発明によれば、ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、ＲＮＡ配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したＲＮＡ配列を抽出し、構造トポロジーとＲＮＡ配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたＲＮＡ配列と構造トポロジーに対応する格子部分に適合度を表示する適合度マトリックスを作成し、適合度マトリックスについて、適合度により構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた構造トポロジーの順番に対応する生成文法により構文解析を行い適合度が最大となる構文解析木を求め、予め定めた条件を満たす適合度を持つ構文解析木に対応する他のＲＮＡ配列を抽出するので、共通の構造を持つＲＮＡ配列を容易に発見することができるＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体を提供することができる。

産業上の利用可能性

以上のように、本発明にかかるＲＮＡ配列解析装置、ＲＮＡ配列解析方法、プログラム、および、記録媒体は、ＲＮＡ二次構造予測、ＲＮＡ配列解析、遺伝子予測、および、これらを利用した創薬等に適している。

請 求 の 範 囲

1． ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、

上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、

上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソート手段と、

上記ソート手段によりソートされた上記構文解析木を上記ＲＮＡ配列の二次構造の候補として出力する出力手段と、

を備えたことを特徴とするＲＮＡ配列解析装置。

2． ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、

上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、

上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力する出力手段と、

を備えたことを特徴とするＲＮＡ配列解析装置。

3． ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、

上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う

適合度計算手段と、

上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出手段と、

上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出手段にて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記ＲＮＡ配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成手段と、

を備えたことを特徴とするＲＮＡ配列解析装置。

４．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、

利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成するＲＮＡ配列作成手段と、

上記ＲＮＡ配列作成手段により作成された上記ＲＮＡ配列に対して上記生成文法を適用して構文解析木を導出する構文解析手段と、

上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、

上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列に対応する上記ＤＮＡ配列部分を遺伝子の候補として予測する遺伝子予測手段と、

を備えたことを特徴とするＲＮＡ配列解析装置。

５．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、

上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、

上記適合度計算手段により計算された上記適合度に基づいて上記ＲＮＡ配列間の類似度を計算する類似度計算手段と、

を備えたことを特徴とするＲＮＡ配列解析装置。

６．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、

上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、

上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出手段と、

上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出手段にて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成手段と、

上記適合度マトリックス作成手段にて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のＲＮＡ配列を抽出する共通構造抽出手段と、

を備えたことを特徴とするＲＮＡ配列解析装置。

７．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソートステップと、

上記ソートステップによりソートされた上記構文解析木を上記ＲＮＡ配列の二次構造の候補として出力する出力ステップと、

を含むことを特徴とするＲＮＡ配列解析方法。

８．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力する出力ステップと、

を含むことを特徴とするＲＮＡ配列解析方法。

９．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を

満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出ステップと、

上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記ＲＮＡ配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成ステップと、

を含むことを特徴とするＲＮＡ配列解析方法。

１０．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成するＲＮＡ配列作成ステップと、

上記ＲＮＡ配列作成ステップにより作成された上記ＲＮＡ配列に対して上記生成文法を適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列に対応する上記ＤＮＡ配列部分を遺伝子の候補として予測する遺伝子予測ステップと、

を含むことを特徴とするＲＮＡ配列解析方法。

１１．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を

行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度に基づいて上記ＲＮＡ配列間の類似度を計算する類似度計算ステップと、

を含むことを特徴とするＲＮＡ配列解析方法。

１２．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出ステップと、

上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成ステップと、

上記適合度マトリックス作成ステップにて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のＲＮＡ配列を抽出する共通構造抽出ステップと、

を含むことを特徴とするＲＮＡ配列解析方法。

１３．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生

成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソートステップと、

上記ソートステップによりソートされた上記構文解析木を上記ＲＮＡ配列の二次構造の候補として出力する出力ステップと、

を含むことを特徴とするＲＮＡ配列解析方法をコンピュータに実行させるプログラム。

１４．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力する出力ステップと、

を含むことを特徴とするＲＮＡ配列解析方法をコンピュータに実行させるプログラム。

１５．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出ステップと、

上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記ＲＮＡ配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成ステップと、

を含むことを特徴とするＲＮＡ配列解析方法をコンピュータに実行させるプログラム。

１６．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成するＲＮＡ配列作成ステップと、

上記ＲＮＡ配列作成ステップにより作成された上記ＲＮＡ配列に対して上記生成文法を適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列に対応する上記ＤＮＡ配列部分を遺伝子の候補として予測する遺伝子予測ステップと、

を含むことを特徴とするＲＮＡ配列解析方法をコンピュータに実行させるプロ

グラム。

１７．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度に基づいて上記ＲＮＡ配列間の類似度を計算する類似度計算ステップと、

を含むことを特徴とするＲＮＡ配列解析方法をコンピュータに実行させるプログラム。

１８．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出ステップと、

上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成ステップと、

上記適合度マトリックス作成ステップにて作成された上記適合度マトリックス

について、上記適合度により上記構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のＲＮＡ配列を抽出する共通構造抽出ステップと、

を含むことを特徴とするＲＮＡ配列解析方法をコンピュータに実行させるプログラム。

１９．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソートステップと、

上記ソートステップによりソートされた上記構文解析木を上記ＲＮＡ配列の二次構造の候補として出力する出力ステップと、

を含むＲＮＡ配列解析方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

２０．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を

行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を上記構造トポロジーの二次構造を持つＲＮＡ配列の候補として出力する出力ステップと、

を含むＲＮＡ配列解析方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

２１．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出ステップと、

上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記ＲＮＡ配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成ステップと、

を含むＲＮＡ配列解析方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

２２．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

利用者が入力したＤＮＡ配列から転写されるＲＮＡ配列を作成するＲＮＡ配列

作成ステップと、

上記ＲＮＡ配列作成ステップにより作成された上記ＲＮＡ配列に対して上記生成文法を適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列に対応する上記ＤＮＡ配列部分を遺伝子の候補として予測する遺伝子予測ステップと、

を含むＲＮＡ配列解析方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

２３．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度に基づいて上記ＲＮＡ配列間の類似度を計算する類似度計算ステップと、

を含むＲＮＡ配列解析方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

２４．　ＲＮＡ二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

ＲＮＡ配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を

行う適合度計算ステップと、

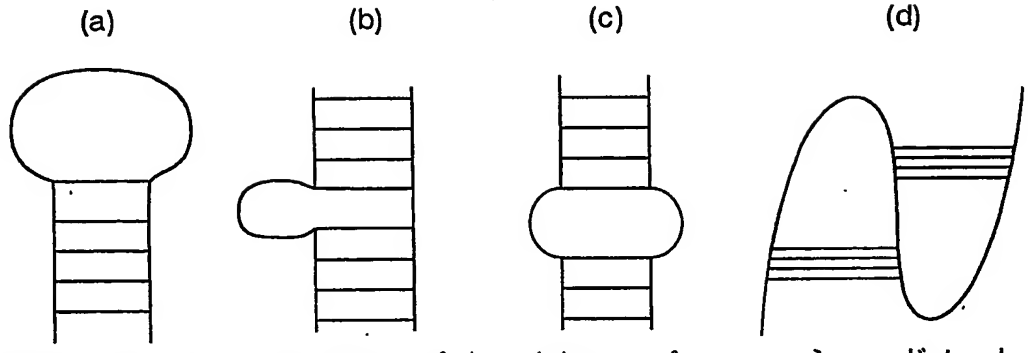
上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記ＲＮＡ配列を抽出する抽出ステップと、

上記構造トポロジーと上記ＲＮＡ配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記ＲＮＡ配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成ステップと、

上記適合度マトリックス作成ステップにて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のＲＮＡ配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のＲＮＡ配列を抽出する共通構造抽出ステップと、

を含むＲＮＡ配列解析方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

第１図



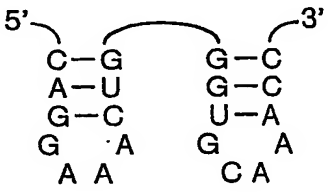
ヘアピンループ (hairpin loop)　バルジループ (bulge loop)　内部ループ (internal loop)　シュードノット (pseudoknots)

第2図

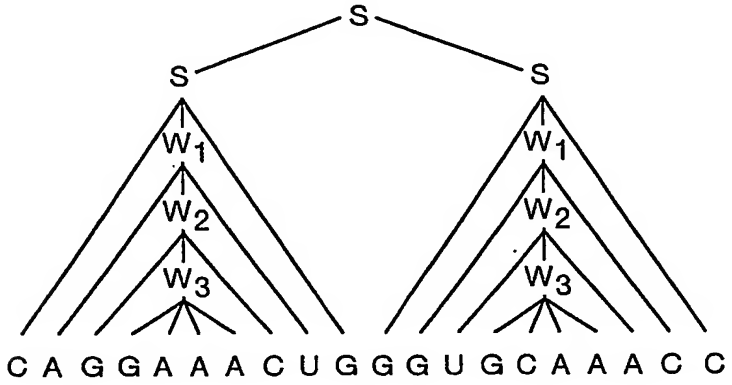
(a)RNA配列

C	A	G	G	A	A	A	C	U	G	G	G	U	G	C	A	A	A	C	C

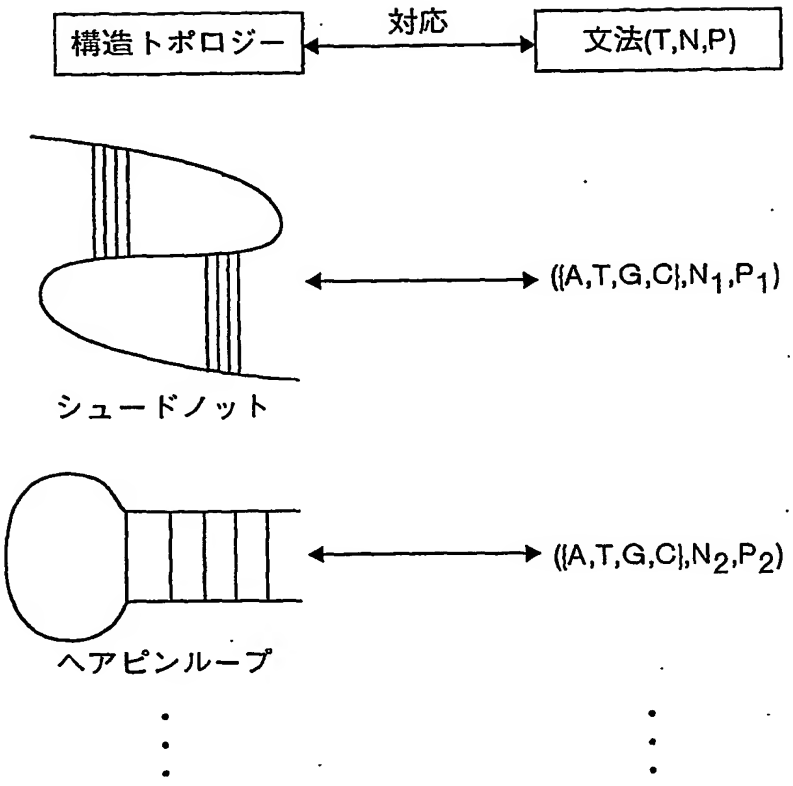
(b)二次構造(ステム構造)



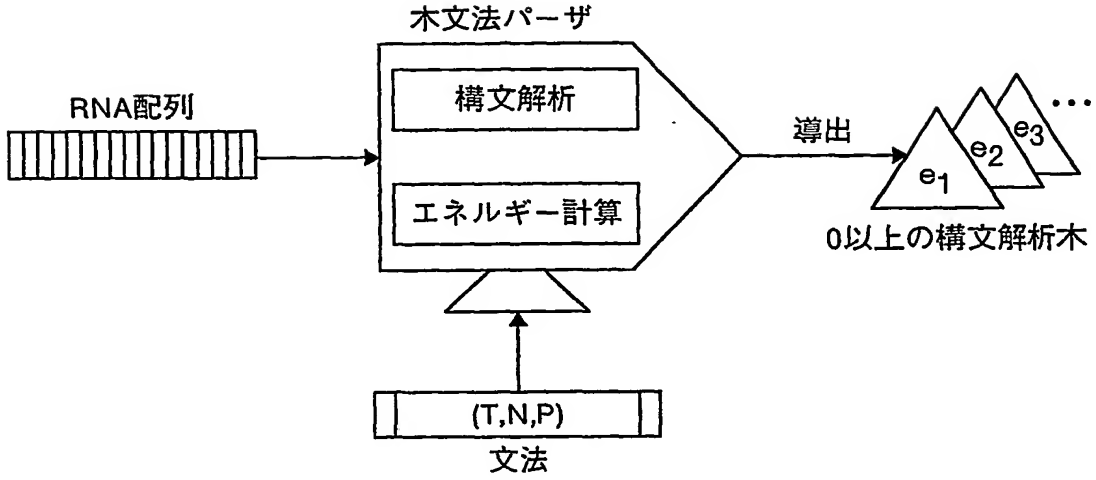
(c)構文解析木



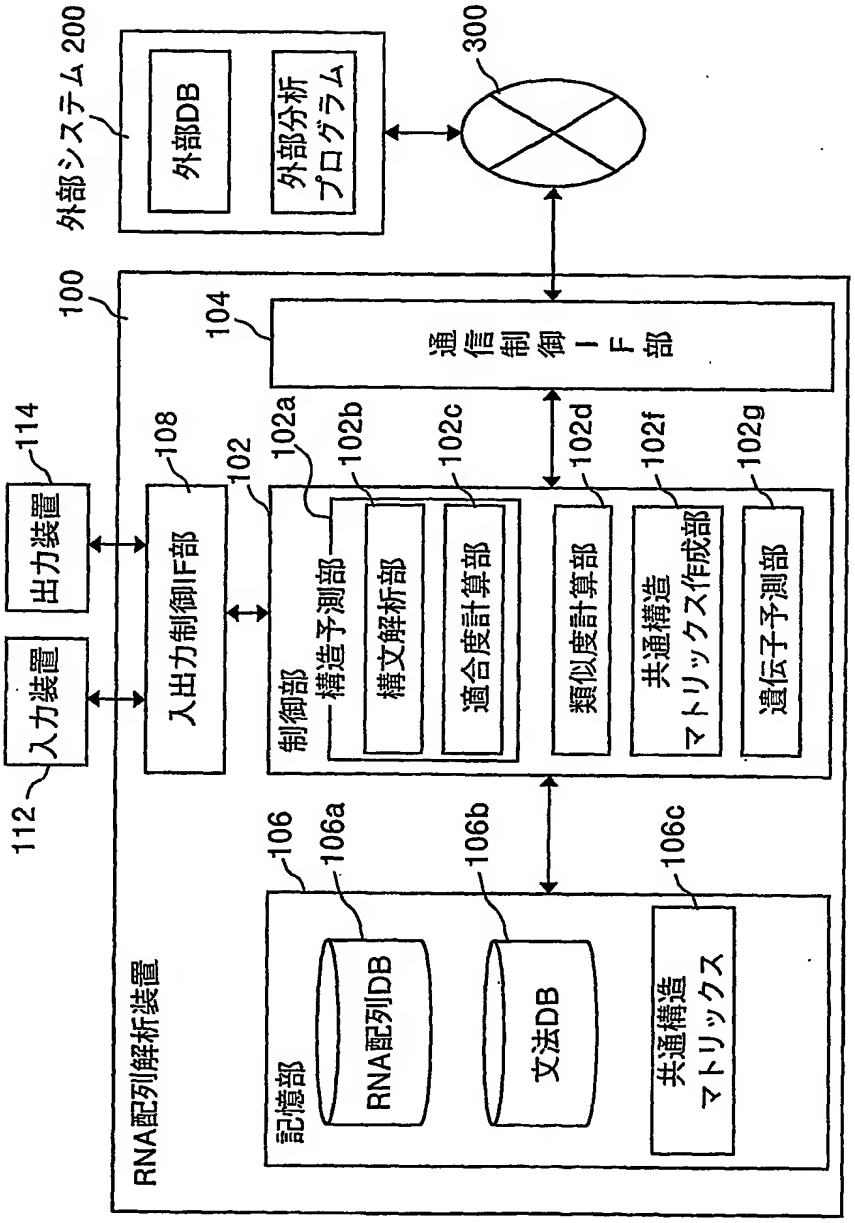
第３図



第4図



第5図

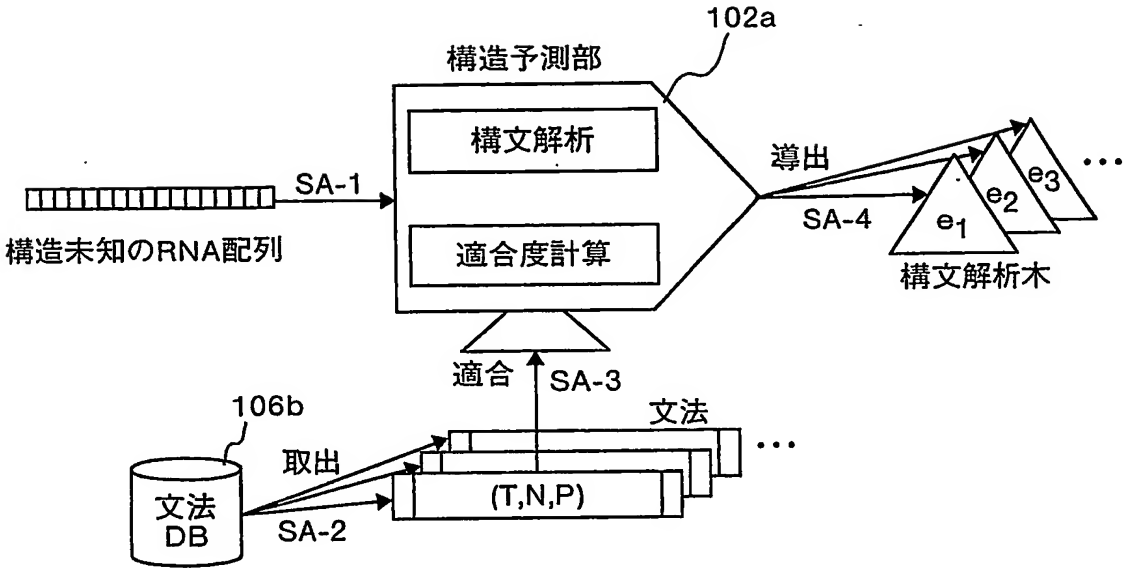

外部システム 200
外部DB
外部分析プログラム
300
100
RNA配列解析装置
104
通信制御I/F部
114
出力装置
108
102
102a
102b
102c
102d
102f
102g
入力装置
入出力制御IF部
制御部
構造予測部
構文解析部
適合度計算部
類似度計算部
共通構造マトリックス作成部
遺伝子予測部
112
106
106a
106b
106c
記憶部
RNA配列DB
文法DB
共通構造マトリックス

第6図

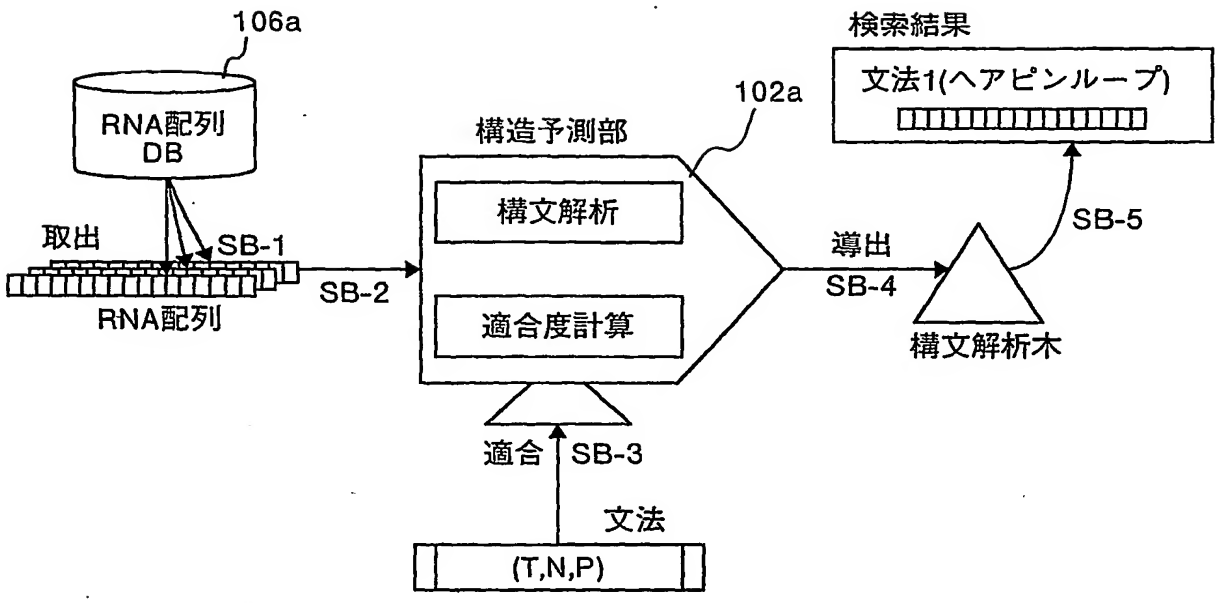
文法データベース 106b

構造トポロジー	文法		
	終端記号T	非終端記号N	生成規則P
シュードノット	{A,T,G,C}	N_1	P_1
ヘアピンループ		N_2	P_2
内側ループ		N_3	P_3
多枝ループ		N_4	P_4
バルジループ		N_5	P_5
⋮		⋮	⋮

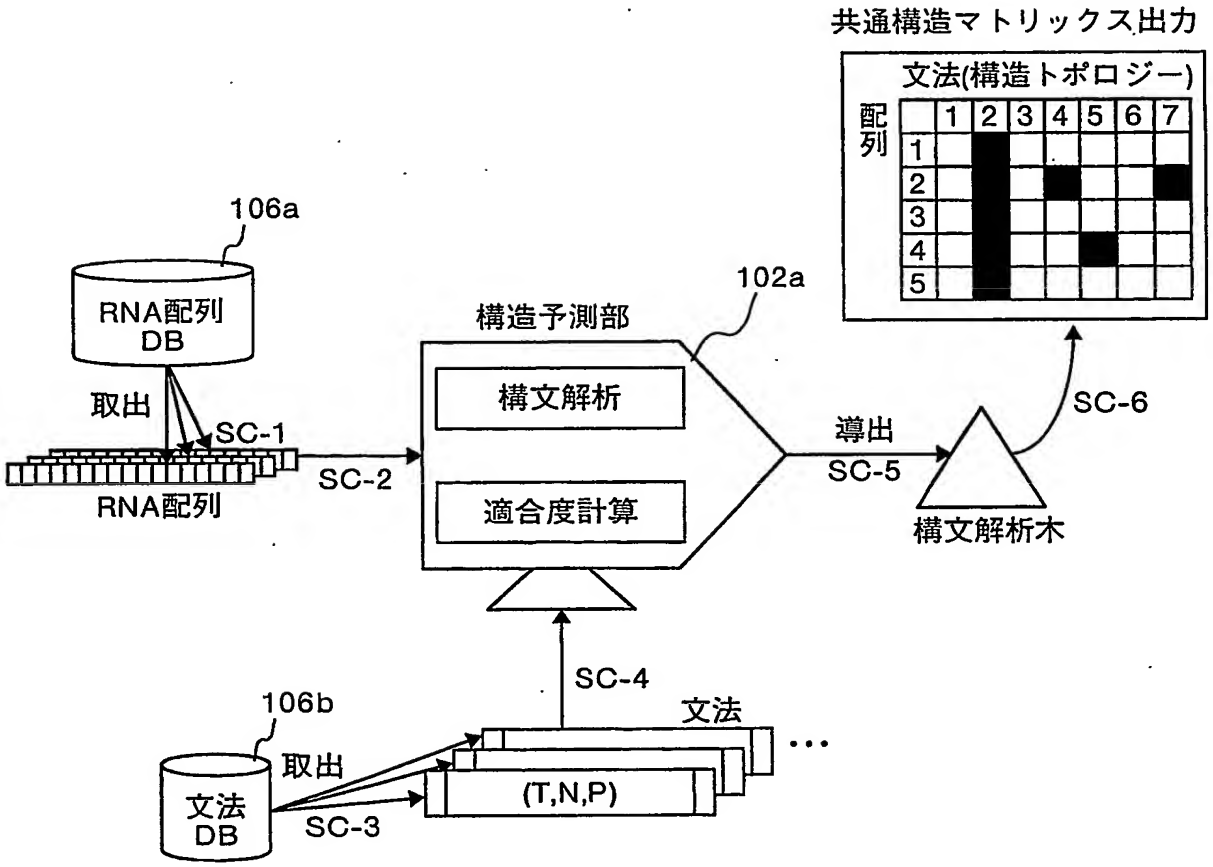
第７図


102a
構造予測部
構文解析
適合度計算
SA-1
構造未知のRNA配列
導出
SA-4
e1
e2
e3
構文解析木
適合
SA-3
106b
文法
取出
文法
DB
SA-2
(T,N,P)

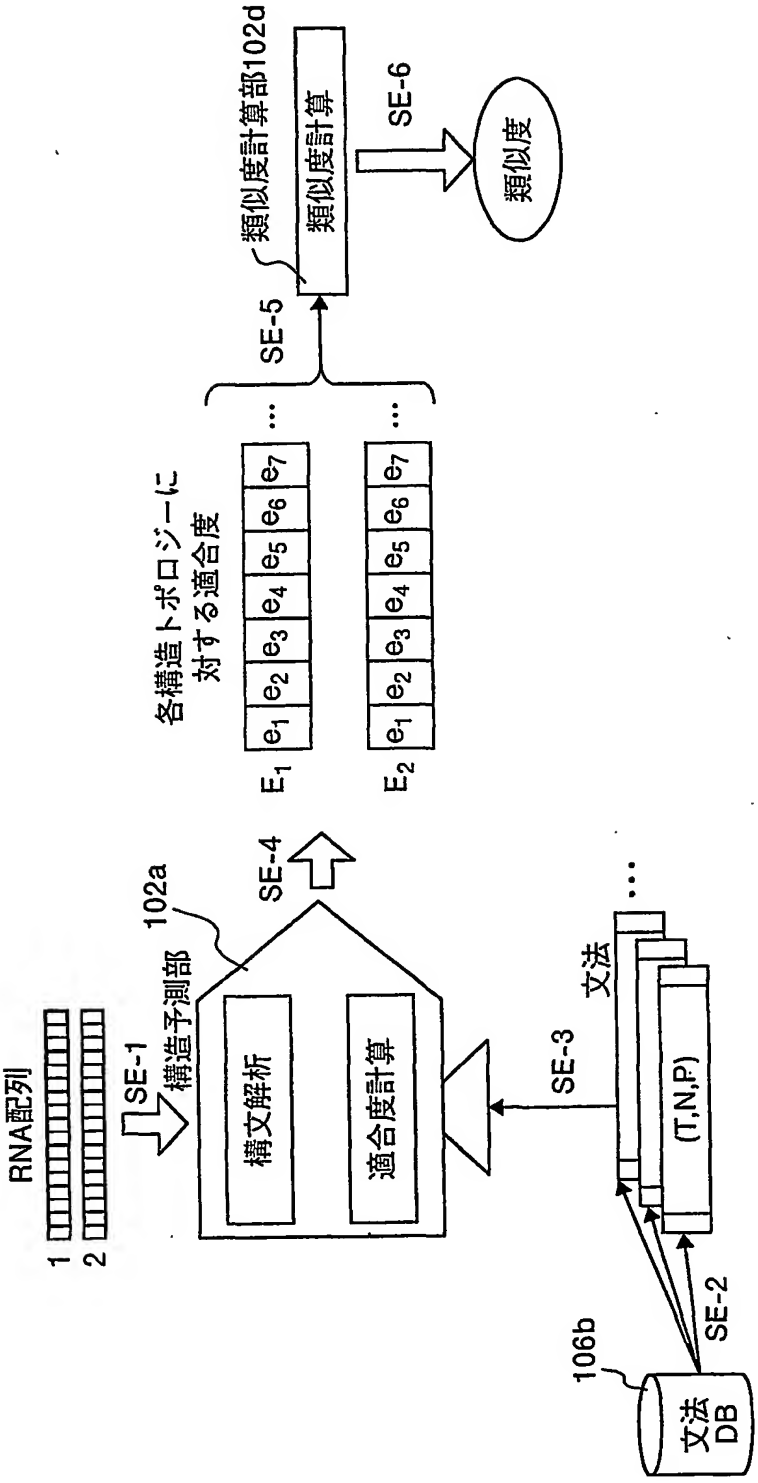
第８図


106a
RNA配列
DB
取出
SB-1
RNA配列
SB-2
構造予測部
102a
構文解析
適合度計算
適合
SB-3
文法
(T,N,P)
導出
SB-4
検索結果
文法1(ヘアピンループ)
SB-5
構文解析木

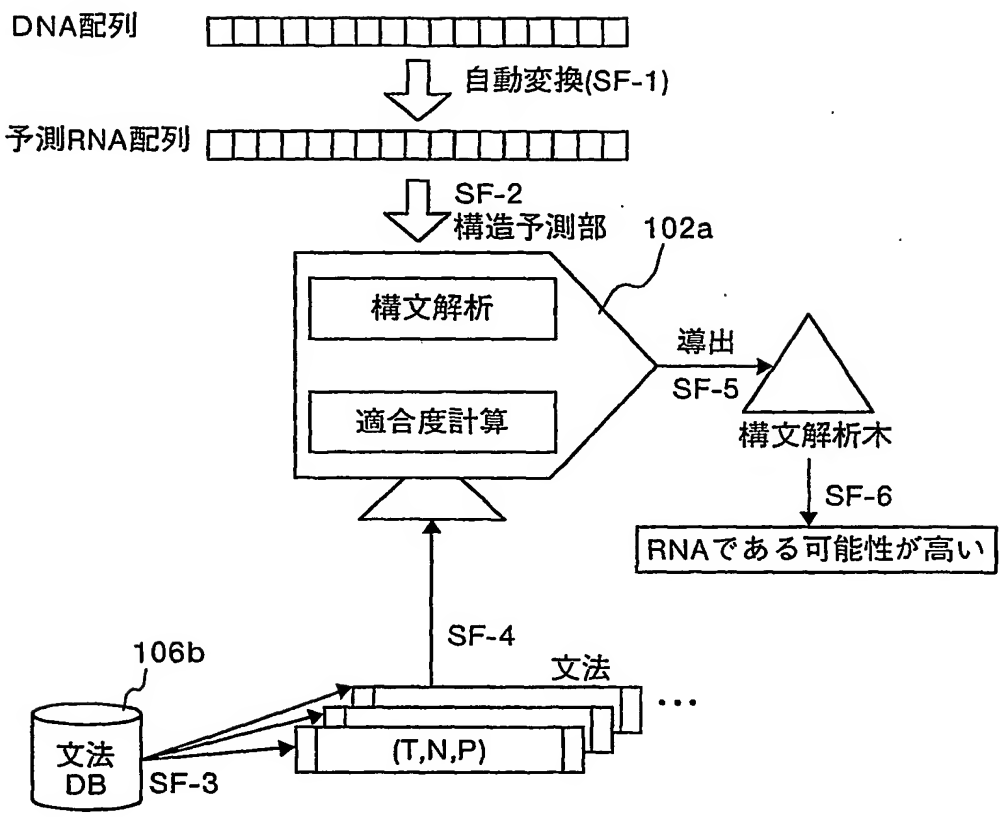
第９図


共通構造マトリックス出力
文法(構造トポロジー)
配列
1 2 3 4 5 6 7
1
2
3
4
5
106a
RNA配列
DB
取出
SC-1
RNA配列
SC-2
構造予測部
102a
構文解析
適合度計算
導出
SC-5
SC-6
構文解析木
SC-4
106b
文法
取出
文法
DB
SC-3
(T,N,P)
...

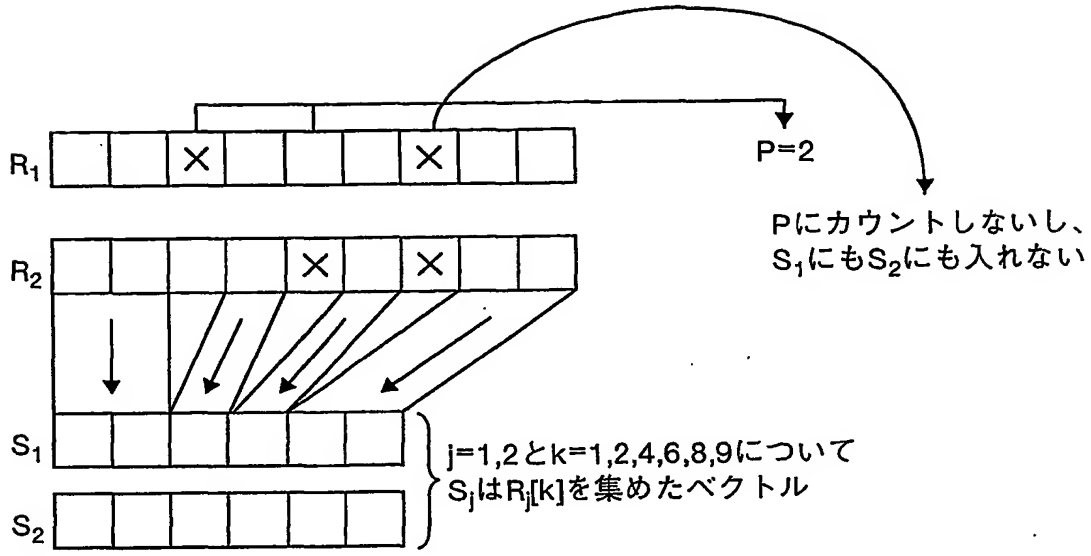
第１０図


RNA配列
1
2
SE-1
構造予測部
102a
構文解析
適合度計算
SE-4
各構造トポロジーに
対する適合度
E_1
E_2
e_1 e_2 e_3 e_4 e_5 e_6 e_7
SE-5
類似度計算部102d
類似度計算
SE-6
類似度
SE-3
文法
(T,N,P)
SE-2
106b
文法
DB

第１１図

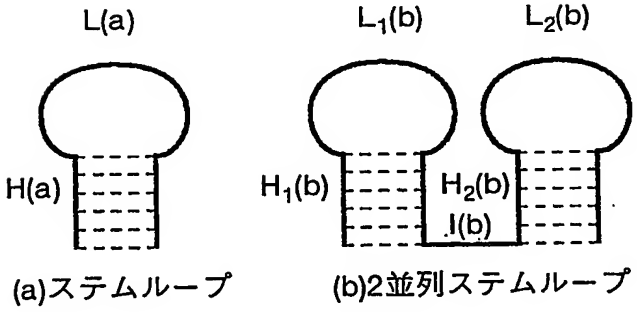

DNA配列
自動変換(SF-1)
予測RNA配列
SF-2
構造予測部
102a
構文解析
適合度計算
導出
SF-5
構文解析木
SF-6
RNAである可能性が高い
SF-4
106b
文法
(T,N,P)
文法
DB
SF-3

第１２図


R1
R2
S1
S2
P=2
Pにカウントしないし、
S1にもS2にも入れない
j=1,2とk=1,2,4,6,8,9について
SjはRj[k]を集めたベクトル

第１３図

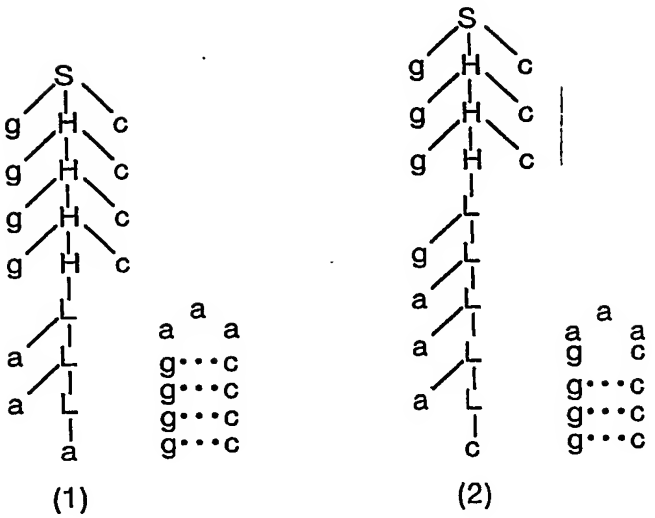
(RNA二次構造トポロジー例)



(a)ステムループ　　(b)2並列ステムループ

第１４図

(s_1の構文解析木と二次構造)



第１５図

(塩基対の自由エネルギー[kcal/mol])

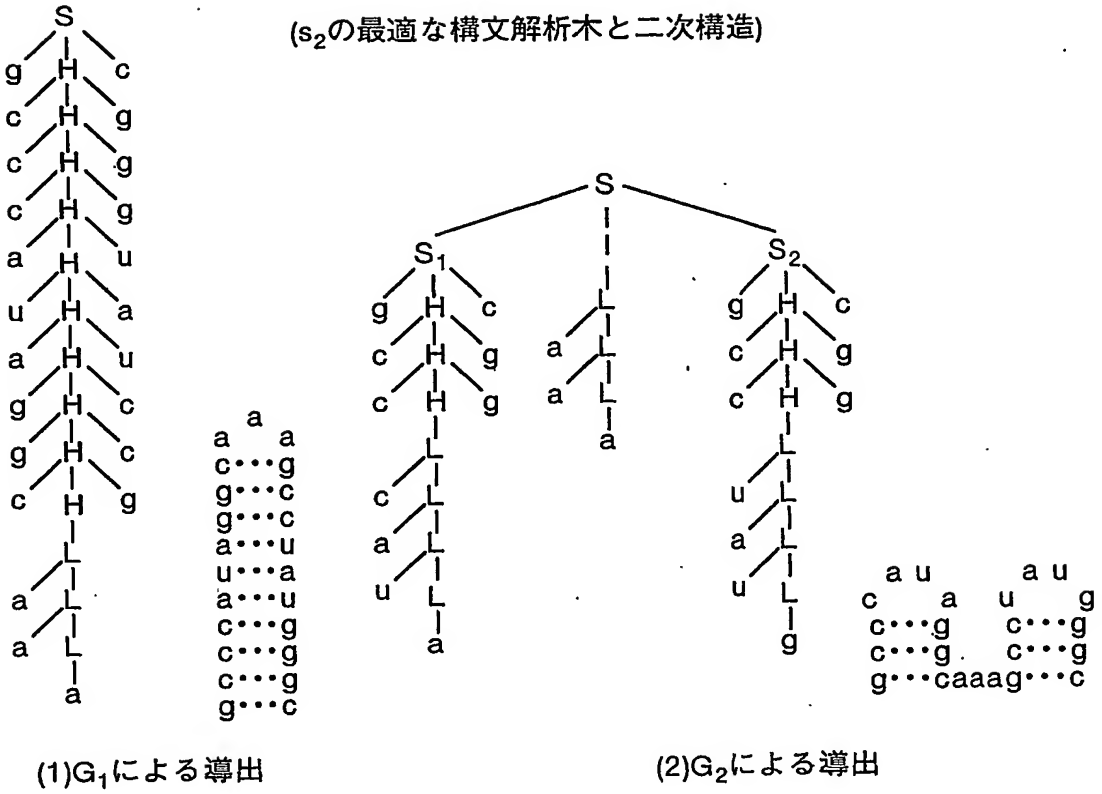
3'側塩基対	5'側塩基対				
	gu	au	ua	cg	gc
gu	-0.5	-0.5	-0.7	-1.5	-1.3
au	-0.5	-0.9	-1.1	-1.8	-2.3
ua	-0.7	-0.9	-0.9	-1.7	-2.1
cg	-1.9	-2.1	-2.3	-2.9	-3.4
gc	-1.5	-1.7	-1.8	-2.0	-2.9

第16図

(ループの自由エネルギー[kcal/mol])

ループ種別	ループサイズ													
	1	2	3	4	5	6	7	8	9	10	12	14	16
バルジループ	3.3	5.2	6.0	6.7	7.4	8.2	9.1	10.0	10.5	11.0	11.8	12.5	13.0
ヘアピンループ	-	-	7.4	5.9	4.4	4.3	4.1	4.1	4.2	4.3	4.9	5.6	6.1
内部ループ	-	0.8	1.3	1.7	2.1	2.5	2.6	2.8	3.1	3.6	4.4	5.1	5.6

第１７図



第１８図

順位	トポロジー	適合度	二次構造
1	ステムループT_1	13.7	図18(1)に示された二次構造
2	２並列ステムループT_2	6.7	図18(2)に示された二次構造

第１９図

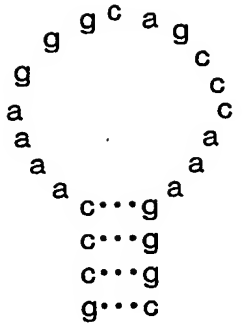
配列	適合度	二次構造
s_2	6.7	図18(2)に示された二次構造

第２０図

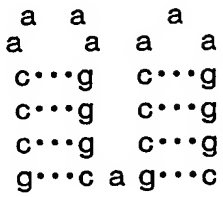
	T_1	T_2
s_1	1.4	-
s_2	13.6	6.7

第２１図

(sの最適な二次構造)



(1)



(2)

第２２図

	T_1	T_2
S	-	5.1

第２３図

配列	適合度	トポロジー(スコア)
s_2	6.7	T_2(5.1)

INTERNATIONAL SEARCH REPORT

International Application No.
PCT/JP03/00011

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06F17/30, C12Q1/68, C12N15/00, G01N33/50

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06F17/30, C12Q1/68, C12N15/00, G01N33/50

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Toroku Jitsuyo Shinan Koho	1994-2003
Kokai Jitsuyo Shinan Koho	1971-2003	Jitsuyo Shinan Toroku Koho	1996-2003

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JICST FILE(JOIS), WPI, INSPEC(DIALOG)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	YOSHIZAWA et al., "RNA Niji Kozo Yosoku Shien System no Kaihatsu to Jisso", The Society for Artificial Intelligence Jinko Chino Kisoron Kenkyukai(Dai 32 Kai) Shiryo, 26 March, 1998 (26.03.98), pages 29 to 35, particularly, page 31	1,2,4,5,7,8,10,11,13,14,16,17,19,20,22,23 3,6,9,12,15,18,21,24
A	UEMURA, Y. et al., Grammatically Modeling and Predicting RNA Secondary Structures. Genome Informatics Series, Proceedings Genome Informatics Workshop, 11 December, 1995 (11.12.95), No.6, pages 67 to 76	1-24

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search 17 February, 2003 (17.02.03)	Date of mailing of the international search report 04 March, 2003 (04.03.03)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (July 1998)

国際調査報告

国際出願番号　PCT／JP03／00011

A.　発明の属する分野の分類（国際特許分類（IPC））

Int. Cl⁷　G06F17／30, C12Q1／68, C12N15／00, G01N33／50

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl⁷　G06F17／30, C12Q1／68, C12N15／00, G01N33／50

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922－1996年
日本国公開実用新案公報	1971－2003年
日本国登録実用新案公報	1994－2003年
日本国実用新案登録公報	1996－2003年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JICSTファイル（JOIS），WPI，INSPEC（DIALOG）

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
X	吉沢 他, RNA二次構造予測支援システムの開発と実装, 人工知能学会人工知能基礎論研究会（第32回）資料, 1998.03.26, p. 29-35 特に、p. 31	1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23
A		3, 6, 9, 12, 15, 18, 21, 24

☒ C欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日 17. 02. 03	国際調査報告の発送日 04.03.03
国際調査機関の名称及びあて先 日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	特許庁審査官（権限のある職員） 高瀬　勤　5M　9069 電話番号　03－3581－1101　内線　3597

様式PCT／ISA／210（第2ページ）（1998年7月）

C（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	UEMURA, Y. et al. Grammatically Modeling and Predicting RNA Secondary Structures. Genome Informatics Series, Proceedings Genome Informatics Workshop, 1995. 12. 11, No. 6, p. 67-76	1-24